Metadata for Information Management in Large-Scale Science

Brian Matthews,
Leader, Scientific Applications Group,
E-Science Centre,
STFC Rutherford Appleton Laboratory

brian.matthews@stfc.ac.uk

Overview

- Science in large facilities
- Supporting large-scale science
- CSMD – a metadata model for science
- ICAT Software Suite
- DataPortal Walkthrough
- Current and Future Directions

Acknowledgements

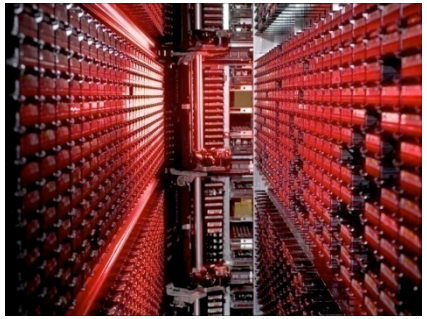
A team effort with many people contributing

Especially: Shoaib Sufi, Kerstin Kleese-van-Dam, Damian Flannery

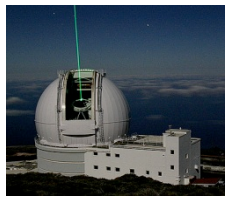
Also: Michael Gleaves, Glen Drinkwater, Louisa Casely-Hayford, Rik Tyer, Ken Shankland, Gordon Brown, Carmine Cioffi, Alun Ashton, Rob Allan, Lisa Blanchard, Juan Bicarregui, Shirley Crompton

Science in Large Facilities

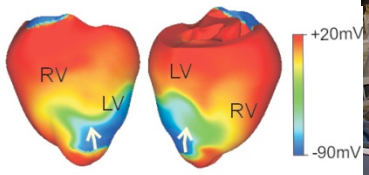
Science and Technology Facilities Council

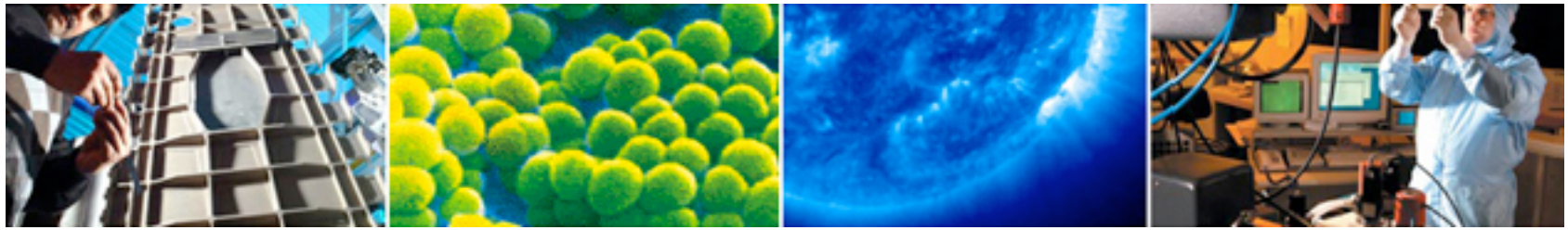


STFC employ more than 2200 staff who are deployed at 7 locations, these are: Swindon where the headquarter is based, the Rutherford Appleton Laboratory, the Daresbury Laboratory, the Chilbolton Observatory, the UK Astronomy Technology Centre in Edinburgh, the Isaac New Group of Telescopes on La Palma; and the Joint Astronomy Centre in Hawaii.



Annually over 15000 visiting scientists.





Research and Science Support at STFC

Deliver world class science

Engender world class science

Communicate world class science

Annually over 15000 visiting Scientists from around the world from both Academia and Industry.

Science & Technology Facilities Council
e-Science

A Multidisciplinary Laboratory

- Spallation Neutron and Muon Source (ISIS)
- Diamond Light Source (DLS)
- Central Laser Facility (CLF)
- Particle Physics (CERN)
- Space Science and Technology (ESA, Observatories, Data Centres)

- Earth Observation
- Atmospheric Science
- Computational Science
- Energy Research
- Information Technology
- Radio Communications
- Surfaces Transforms and Interfaces
- Microstructures
- Molecular Spectroscopy

Large-Scale Science

Science & Technology Facilities Council
e-Science

STFC e-Science Centre

Exploit e-Science technologies throughout STFC's programmes, the research communities they support and the national science and engineering base.

- Grid, HPC, Data storage, Libraries, Data Management, Visualisation
- R&D programme
- c.80 staff

http://www.e-science.clrc.ac.uk/

Supporting Large Scale Science

The Problem

Scientific institutions generate vast quantities of data

- CLRC - ISIS, SRS, Space Science, Particle Physics, Computational Science, ...

More data coming on stream all the time:

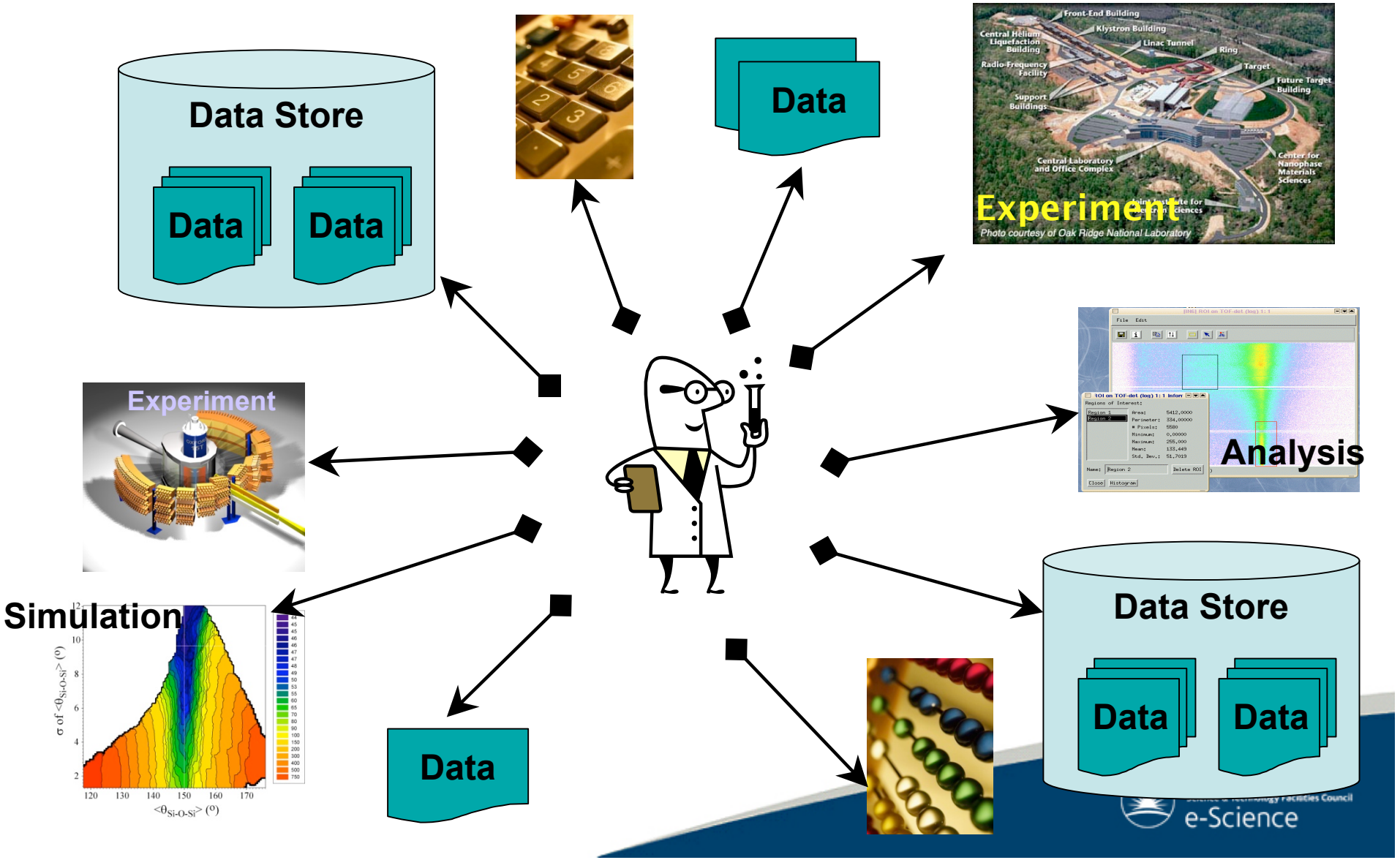
- CERN-LHC, Diamond, XFEL, ...

Very good at handling large amounts of data

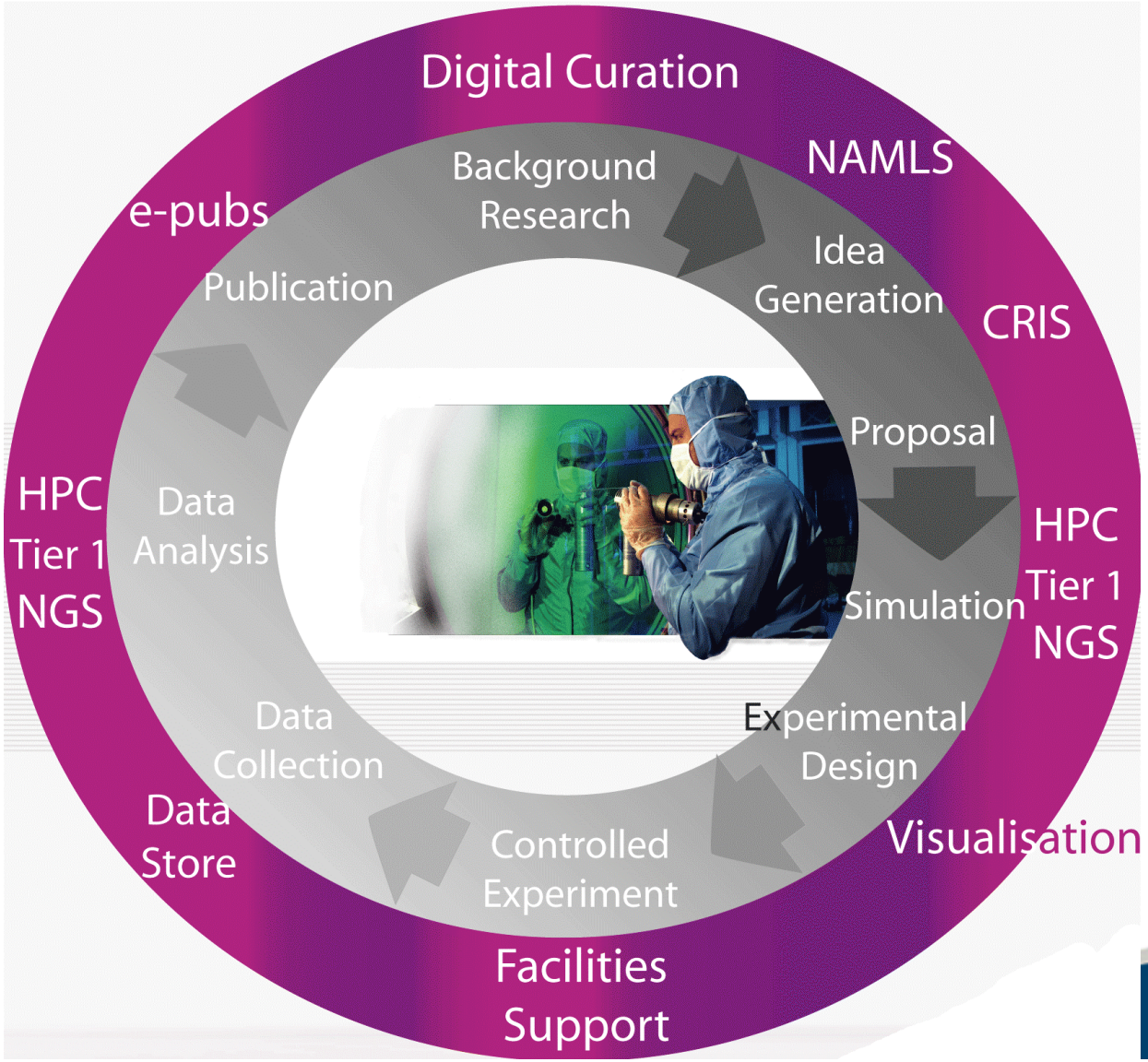
Diverse approaches to organising and distributing it.

Many tools used together in the lifecycle of research

Complexity of large-scale research



The Research Lifecycle



E-Science: providing the infrastructure for the research lifecycle

Science & Technology Facilities Council
e-Science

How do we speed up this research lifecycle?

By speeding up the cycle we can increase the volume of good science

- Make a better return from the investment in science
- Make breakthroughs in science earlier

Do this via:

- Integration
  - Support the whole lifecycle
- Interoperability
  - Support across lifecycles

User Scenarios

Lecturer:

- This *published* study would be a good example for teaching, is the raw data publicly available?

Researcher:

- This is an interesting *paper* - can I check the data?

Experiment Proposer:

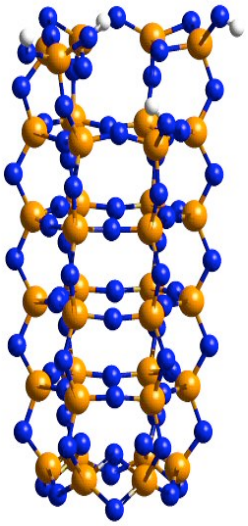
- Have there been any neutron or X-Ray studies of this molecule at 100 K? What *reports* and *papers* have been published on them?

Instrument Scientist:

- The instrument seems a bit unstable recently, fetch me the results of all calibration runs from the last 3 months?

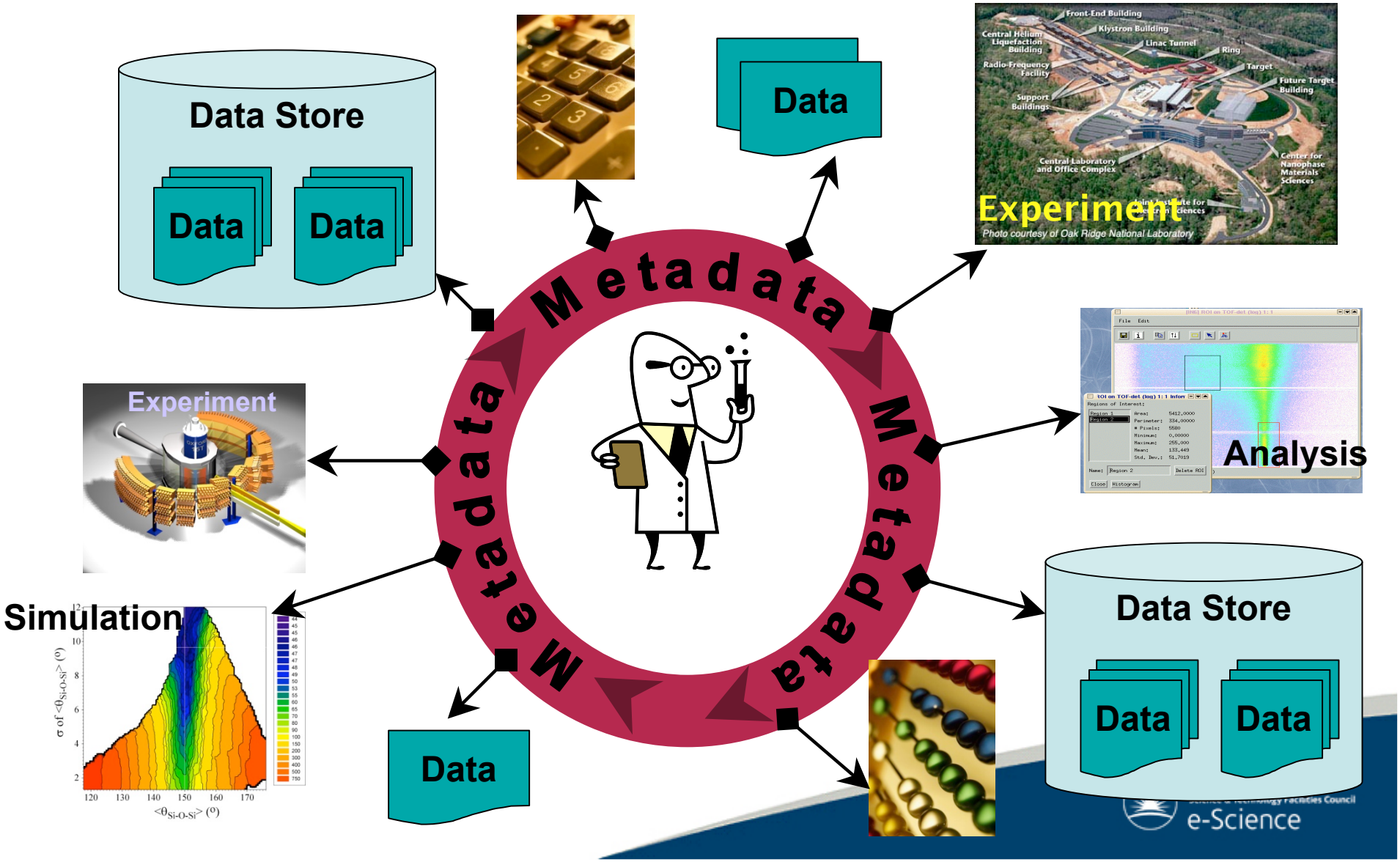
What we aim to achieve with the e-Infrastructure

Enabling users to get rapid access to their current and past data, related experiments, publications etc., leading to improved analysis through more complete information.

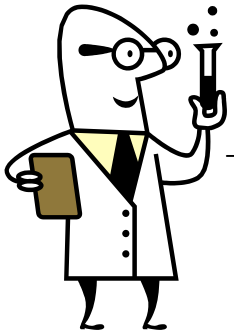


Creating a powerful, long lasting scientific knowledge resource.

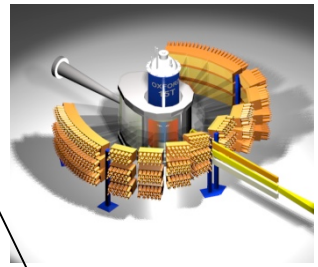
Integration via Metadata



Integrated e-Infrastructure

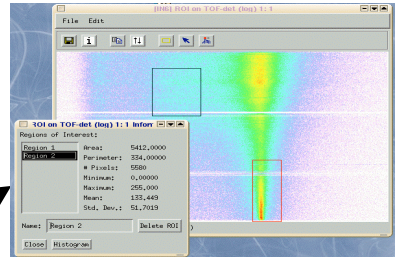


Experiment



Data Acquisition System

Data Analysis

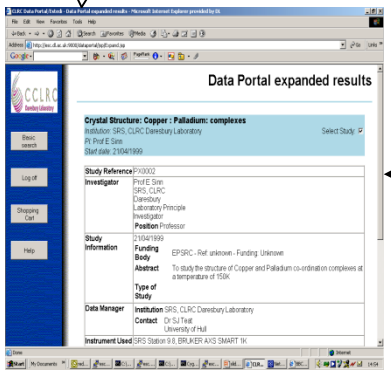


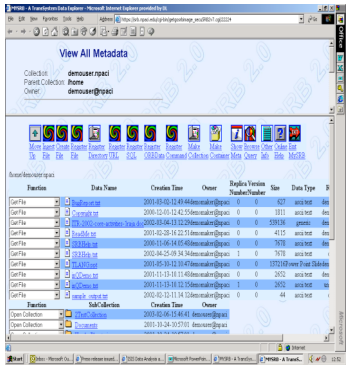
Proposal

All Data and Metadata Capture is automated.

Publication

Information





Proposal System

Metadata Catalogue

Secure Storage

E-Pubs

Science & Technology Facilities Council
e-Science

The ICAT Concept

A method of access to the STFC data resources

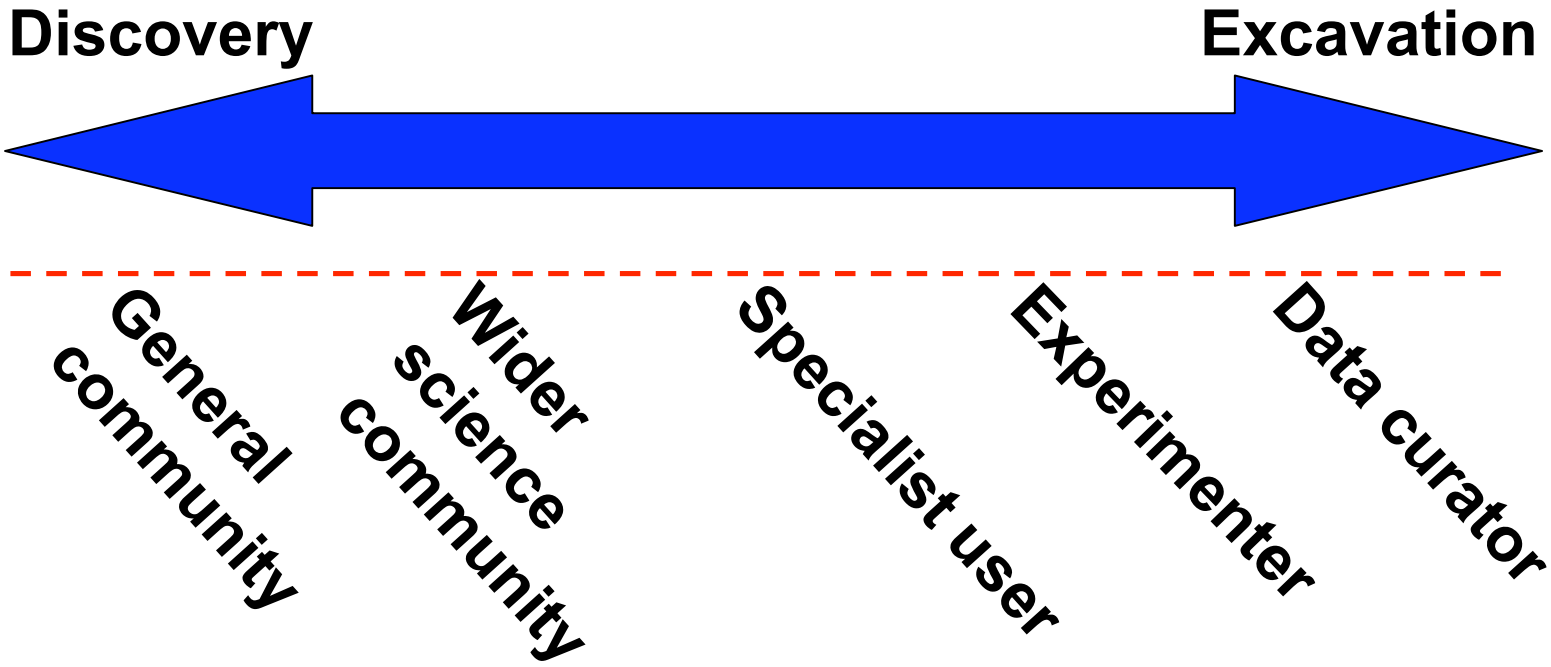
How about a system that would give access to all of it independent of where it was produced?

Encompasses a wide range of data holdings
- Describes what data is available from the facilities
- Links to the data held at the facility
- Different archiving methods

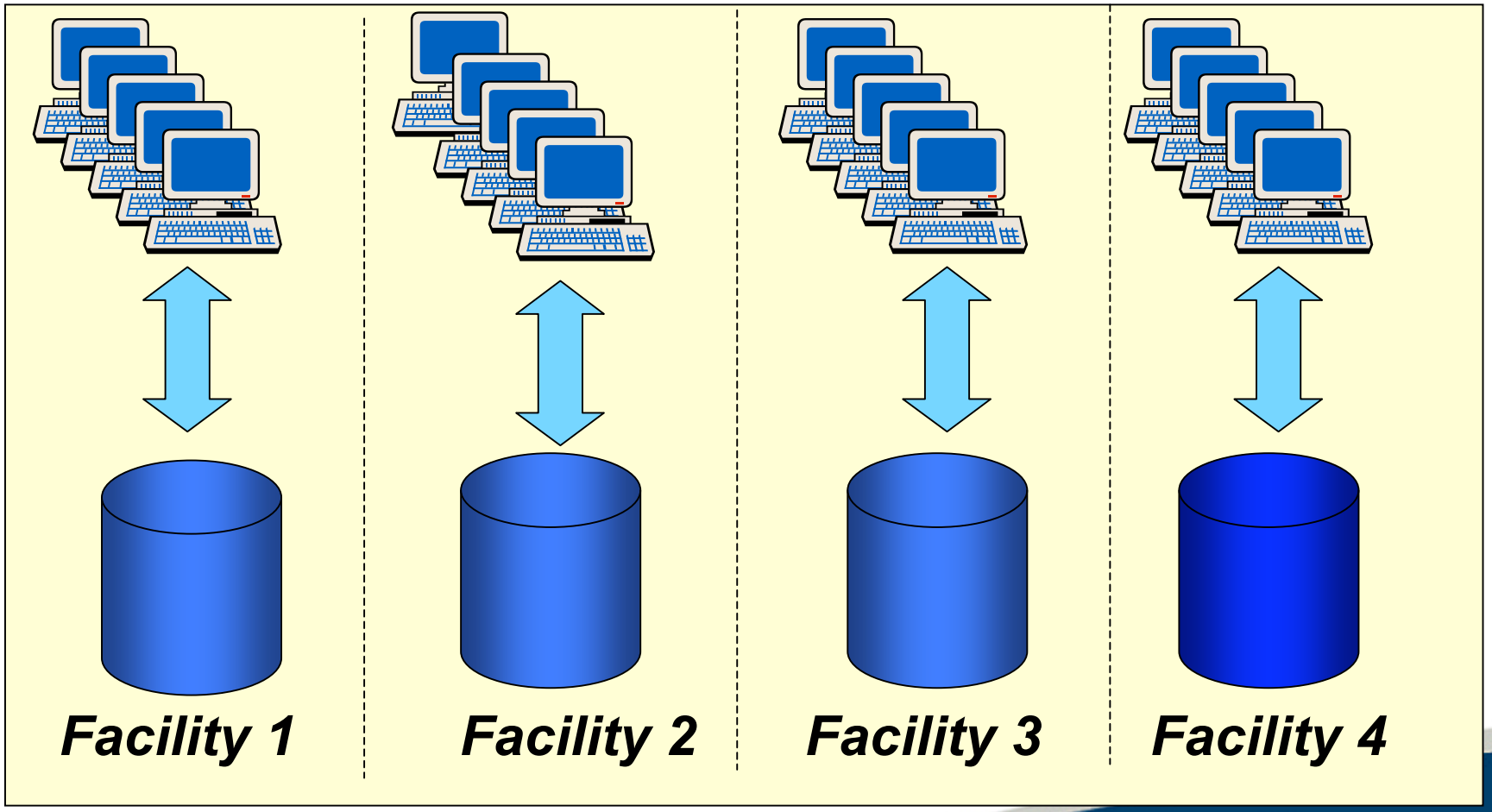
Caters for a wide range of users
- general community → data curators

Supports a wide range of queries
- Ultimately employing data mining, thesauri, ….

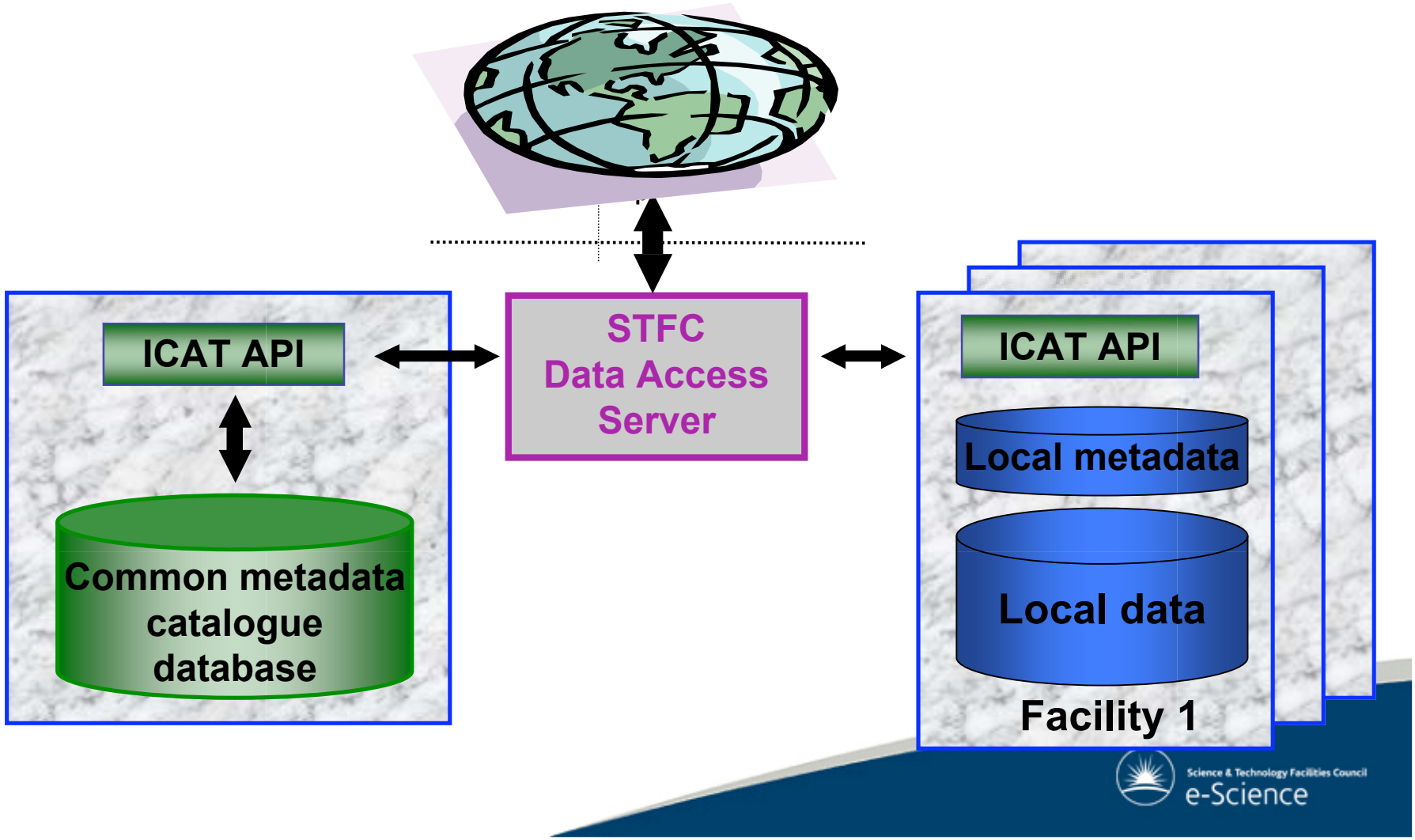
Combine Diverse Users & Searches ...



… with Distributed Data Silos….



…using a central common metadata index …



… and a Web based interface

Exploit the existing Web infrastructure.

- Use Web Technologies (XML/RDF);
- rapidly disseminated;
- widely accessible;
  - *database and user platform independent*
- also being deployed onto the GRID

Every user who needs to can get to the information.

Core Scientific Metadata Model

(CSMD)

Model Motivation (1)

Most Scientists think in terms of Studies during which they perform a number of investigations e.g. experiments, observations, measurements and simulations.

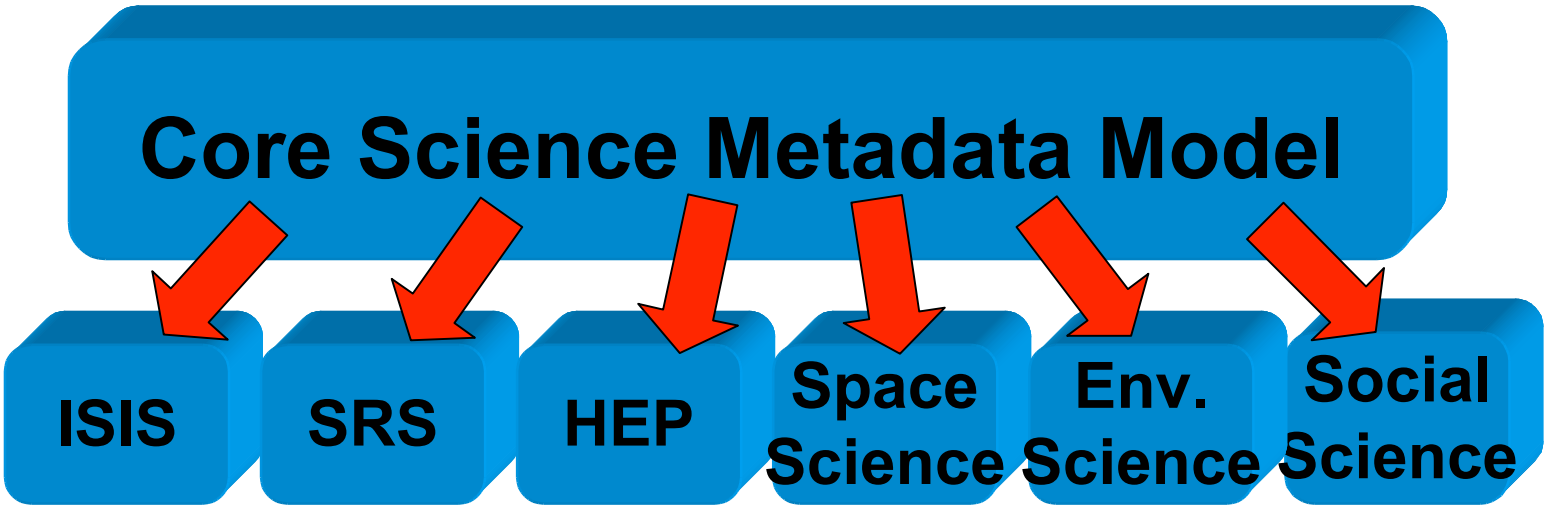
Results from these investigations usually run through different stages: raw data, analysed or derived data and end results. Data should be grouped accordingly.

Metadata and Software (e.g. STFC DataPortal) should allow the user to search for interesting data.

Not all information captured in specific metadata schemas e.g. CML, would be used to search for this data or distinguish one data set from another, give possibility to select special parameter.

Metadata

A generic metadata model for all scientific applications with Specialisation for each domain



Can answer questions across domains

Can answer questions about specific domains

What influenced CSMD?

Work started in 2001:

- CIP from Earth Observation
- DDI from Social Sciences
- DublinCore from the Library community
  - Publication only metadata
- CERIF – research project information
- XSIL as used on LIGO
  - Low level 'Scientific Data Objects' focus
- CERA from the MPIM
  - A bit specific to Earth Sciences but closer

… hence the need to develop out own General Model

Model Motivation (2)

A common general format/standard for Scientific Studies and data holdings metadata did not exist

By proposing Model and Implementation:

- **A specification for the types of metadata which should be captured during Scientific Studies**
- **Ease citation, collaboration, exploitation and Integration**
- **Allow easy Integration of distributed heterogeneous metadata systems into a homogeneous (virtual) Platform**

Therefore – The Common Scientific Metadata Model (CSMD) developed.

Some Model Aims

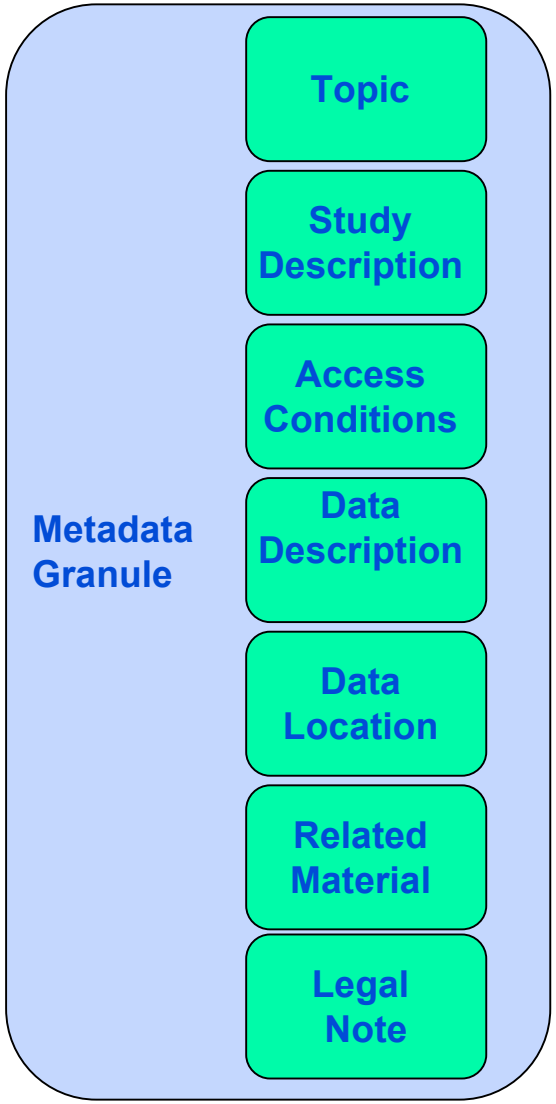
Abstract class orientated description of the types of metadata that should be captured by Scientific Studies

Create a denominator for Scientific Study metadata which form a specification

Provide representations in XML, RDF etc.

Form the basis of a Database Schema representation.

Core Scientific Metadata Model



Keywords **providing a index on what the study is about.**

Provenance **about what the study is, who did it and when.**

Conditions of use **providing information on who and how the data can be accessed.**

Detailed description **of the organisation of the data into datasets and files.**

Locations **providing a navigational aid to where the data on the study can be found.**

References **into the literature and community providing context about the study.**

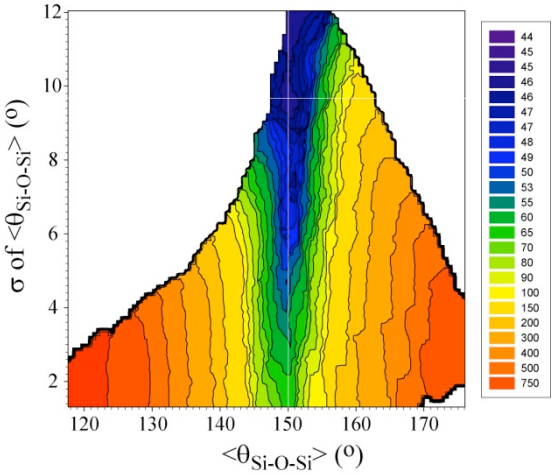
Copyright, patents and conditions of use etc relating to the study and the data in the study

.

Core Scientific Metadata Model (2)

Detailed Information about Instrument and Experiment such as:

- **Sample Information and Parameter**
- **Experimental Station and Set Up**
- **Environmental Parameters**
- **Key Parameters from the Data**
- **Keywords and Classifications**



Model Breakdown: Provenance

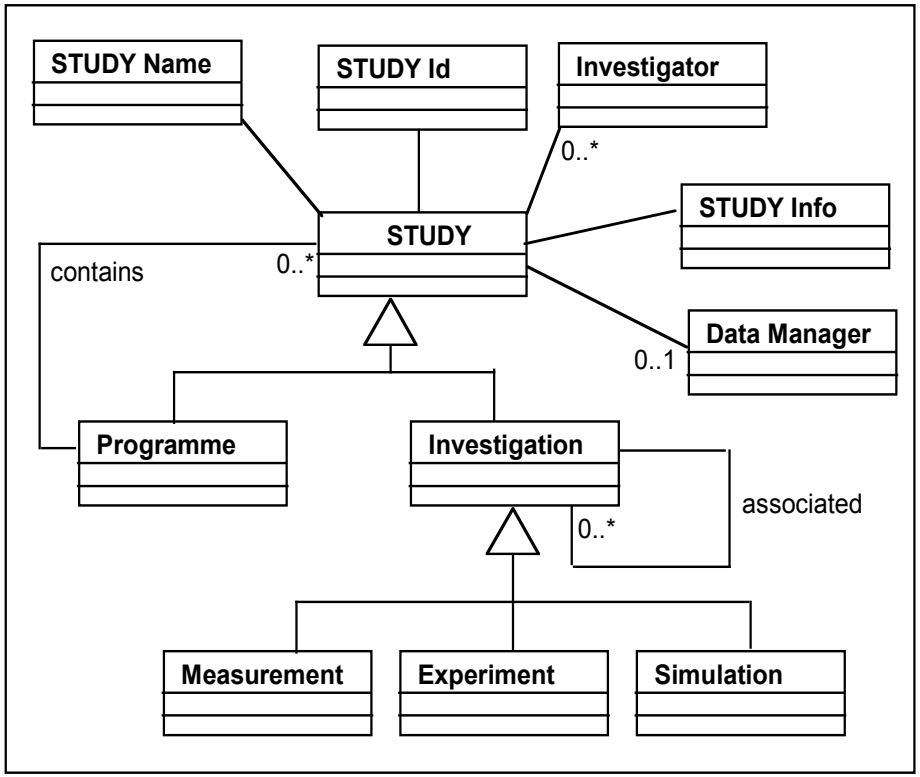
The Study is the basic unit for a scientific activity.

The Study contains the following metadata:

- The Study Name
- The Study Institution
- The Investigator
- Extended Study Information
  - Abstract
  - Funding
  - Start and End times

Can be further divided into:

- Programmes: for connected studies.
- Investigations: for a single measurement, experiment or simulation.



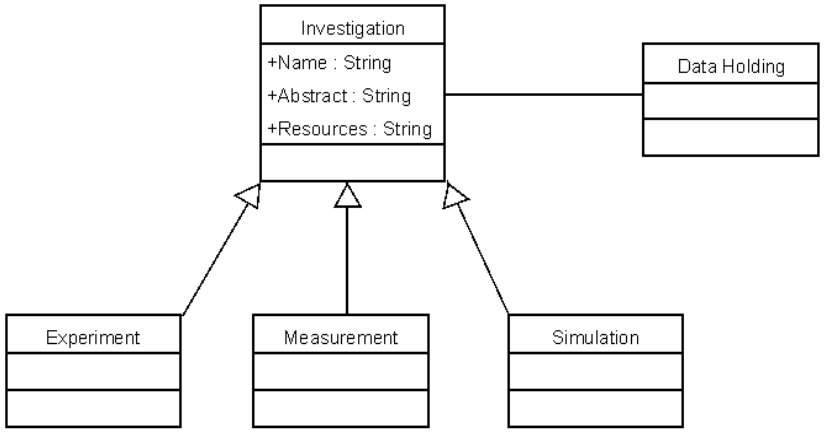
Investigations

A Study can have more than one investigation;

- experiment,
- simulation,
- measurements etc.

investigations contain:

- Name
- Investigation Type
- Abstract
- Resource
- Link to DataHolding



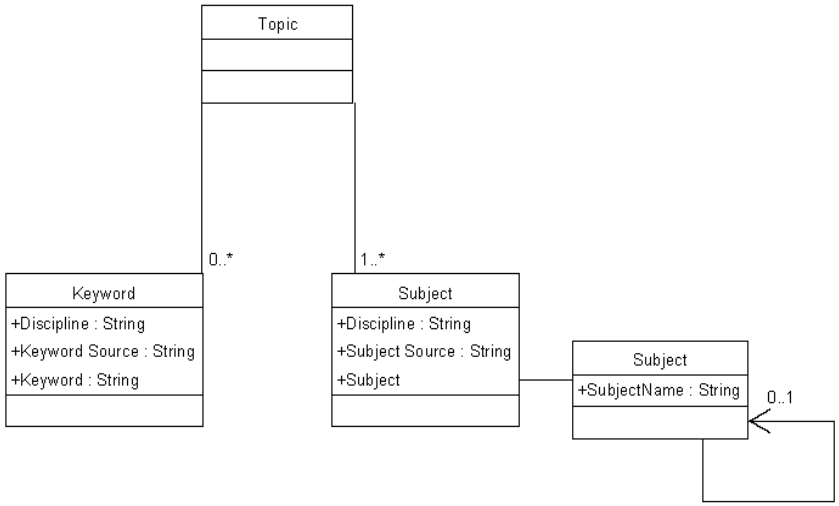
Topic (for indexing)

Keywords

- Discipline (i.e. domain)
- Keyword Source (e.g. domain dictionary)
- Keyword

Subjects

- Discipline
- Subject Source (e.g. domain taxonomy)
- Subject



Access Condition & Related Material

Access Conditions

- Contains a list of users or groups who are allowed access to the metadata and data,
- or a pointer to an access control system which contains such data for this study

Related Material

- One or many links and or textual descriptions of material related to this study
  - e.g. earlier studies or parallel studies
  - Cited publications

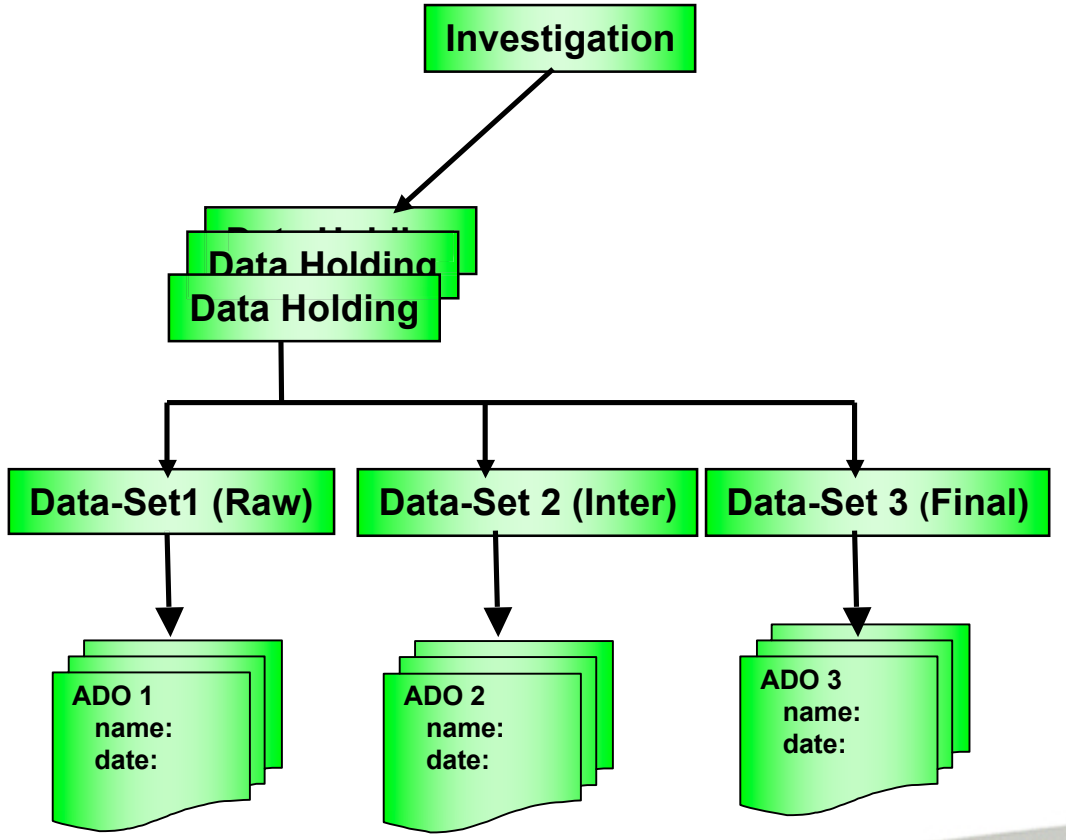
Hierarchy of Data Holdings

Data holdings are associated *w*ith investigations.

These are themselves arranged in a hierarchy:

- Data Collections
- Atomic Data Objects (e.g. files), with links between them

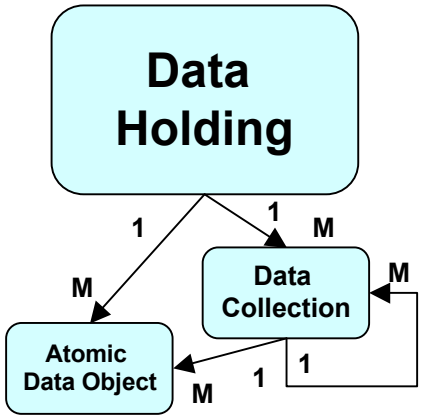
Logical organisation – identity separated from location.



Data

Data Description holds a logical description of the Study's data:

- Data Name
- Type of Data
- Status
- Data Topic
- Parameters
- Related Data Ref
- Relation type (e.g. derived)



Data Location contains the link between logical name (e.g. URI's) and physical URLs

- Data Name
- Locator(s)
- In the case of Atomic Data Objects
  - these can refer to files
  - or as named Selects on a database – i.e. virtual data objects)

More on Parameters

Parameters contain a lot of information about the atomic data objects (ADO) and collections

A collection/ADO can have many parameter entries, each parameter entry contains:

Parameter derivation (e.g. measured/fixed)

- The value
- The units
- Range
- Error margin

Parameter aggregation is also supported

Cardinality Issues

The model recommends a certain cardinality of elements

Certain metadata components are necessary for one to have an instance of the implemented model – treating everything as optional is not acceptable

It is thought implementations may modify this more to their needs – model attempts to remain ideal (i.e. most common Cardinality)

Enumeration Issues

Enumerations (or controlled vocabularies) e.g. types of investigator, types of institutions; these are distinct from the model e.g. as taxonomies are.

However they are necessary for the model to work so implementations e.g. STFC DataPortal XML implementation of the model propose some enumerations for common things

Recognised and relevant controlled vocabularies are hoped to be used by implementations where they are available

Developing a RDF based controlled vocabulary:

See later

CSMD Used on DataPortal

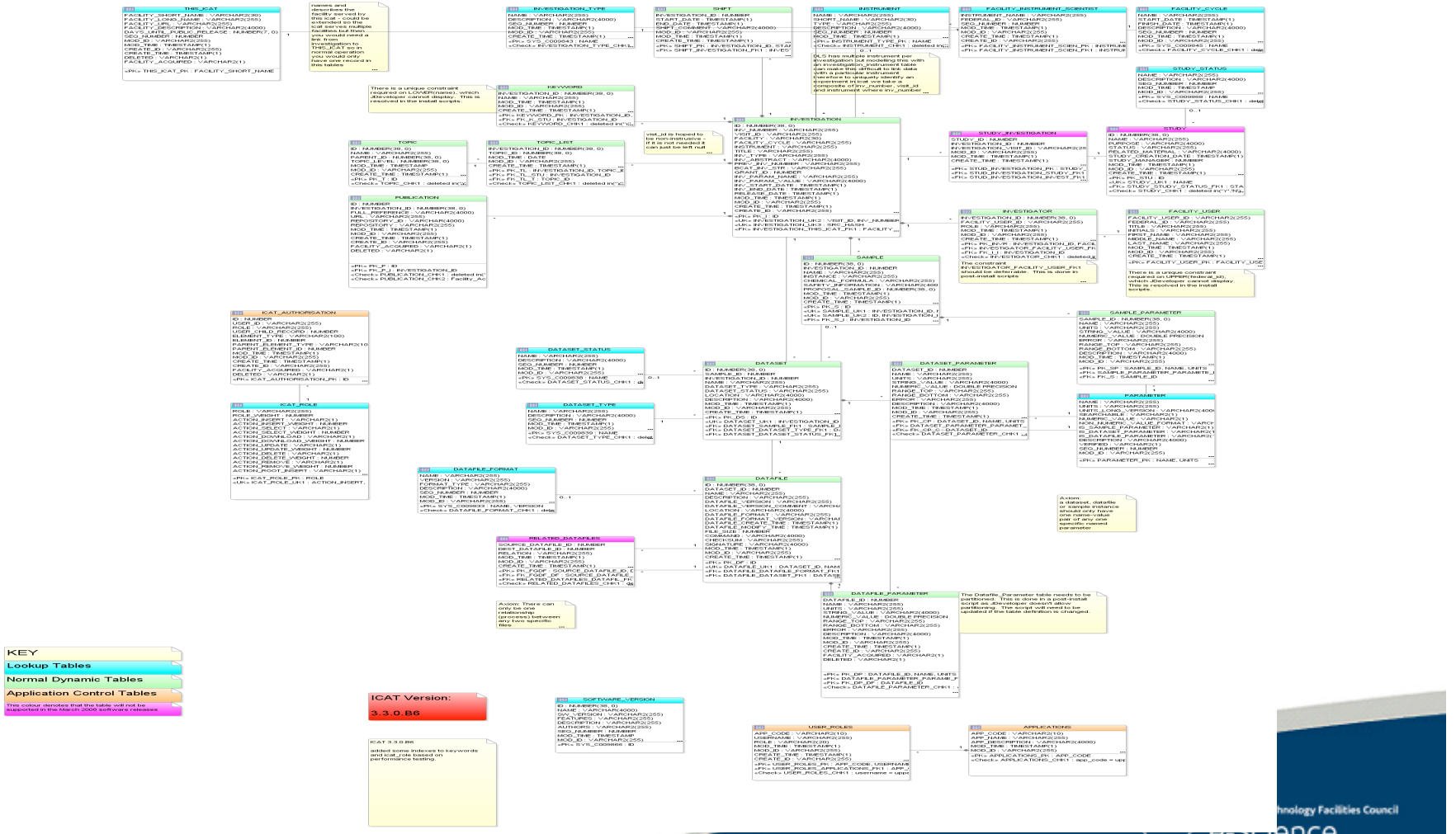
- Implementation used as Data Interface for DataPortal
- Single view of heterogeneous systems/schemas
- Acts as a stress test o the model
  - Limitations feed into Model Requirements
  - New requirements feed back into implementation

```
<MetadataRecord metadataID="S1">
        <Topic>
                <Discipline>Inorganic Chemistry</Discipline>
                <Subject>Copper</Subject>
                <Subject>Crystal</Subject>
                <Subject>Crystal Structure</Subject>
                <Subject>Inorganic Chemistry</Subject>
                <Subject>Metals</Subject>
                <Subject>Palladium</Subject>
        </Topic>
        <Experiment>
                <StudyName>Crystal Structure: Copper : Palladiu
                <StudyID studyid="PX0002">
                        <Institution institutionID="INST1" inst
                </StudyID>
                <Investigator>
                        <Name>
                                <Surname>Sinn</Surname>
                                <Initials>E</Initials>
                                <PersonTitle>Prof</PersonTitle>
                        </Name>
                        <Status>Professor</Status>
                        <Institution institutionID="INST1" inst
                        <Role>Principle Investigator</Role>
```

Metadata example

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CLRCMetadata SYSTEM "clrcmetadata.dtd">
<CLRCMetadata><MetadataRecord metadataID="N000001">
        <Topic>
                <Discipline>Chemistry</Discipline>
                <Subject>Crystal Structure</Subject>
                <Subject>Copper</Subject>...
        <Experiment>
                <StudyName>Crystal Structure: Copper : Palladium:  :complex: 150K ...
                <Investigator><Name><Surname>Porter...<Institution>University of  Peebles
...
                <Funding>EPSRC ...
                <TimePeriod><StartDate><Date>21/04/1999....
                <Purpose><Abstract>
                    To study the structure of Copper and Palladium co-ordination complexes at a 150K.
                <DataManager><Name><Surname>Teat...
                <Instrument>SRS Station 9.8, BRUKER AXS SMART 1K...
                <Condition>...Wavelength...<Units>Angstrom...<ParamValue>0.6890...
                <Condition>…Crystal-to-detector distance<Units>cm...<ParamValue>5.00...
        <AccessConditions>The user has to be one of: Prof. F. Porter….
```

ICAT 3.3 Database Schema



Conformance Level

For a complete metadata study-dataset record a large amount of metadata has to be stored/processed

So it's useful to have conformance levels

Model uses 5 levels

Each level specifies more metadata (and Indexing information) should be held

Benefit of conformance levels; the higher the level of conformance to the CSMD the richer the clients that operate on the data can be

- e.g. identifying datasets and atomic data objects which link directly to keywords/taxonomies and not just studies

Level 1

Type of Information captured:

- Study and Investigation metadata with indexing at the Study level

Level 1 metadata is similar to library/publication style metadata (e.g. DublinCore)

Level 2

Type of Information captured:

- Level 1 + DataHolding metadata (i.e. DataSets and DataObjects)

Level 3

Type of Information captured:

- Level 2 + related material, Access condition, indexing to data collection levels

Level 4

Type of Information captured:

- Level 3 + indexing to data object level and data object parameter information

Level 5

Type of Information captured:

- All metadata components are filled as L4 + funding, resources used, facilities used etc

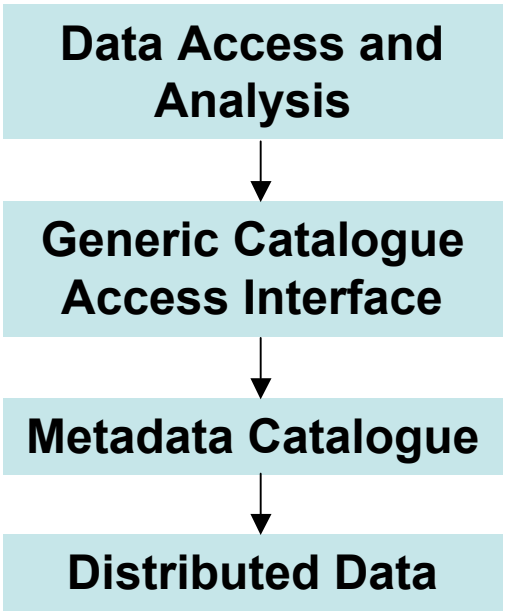
The current DataPortal uses somewhere between L4 and L5 – the new systems designed with CSMD conforms to L4+

ICAT Software Suite

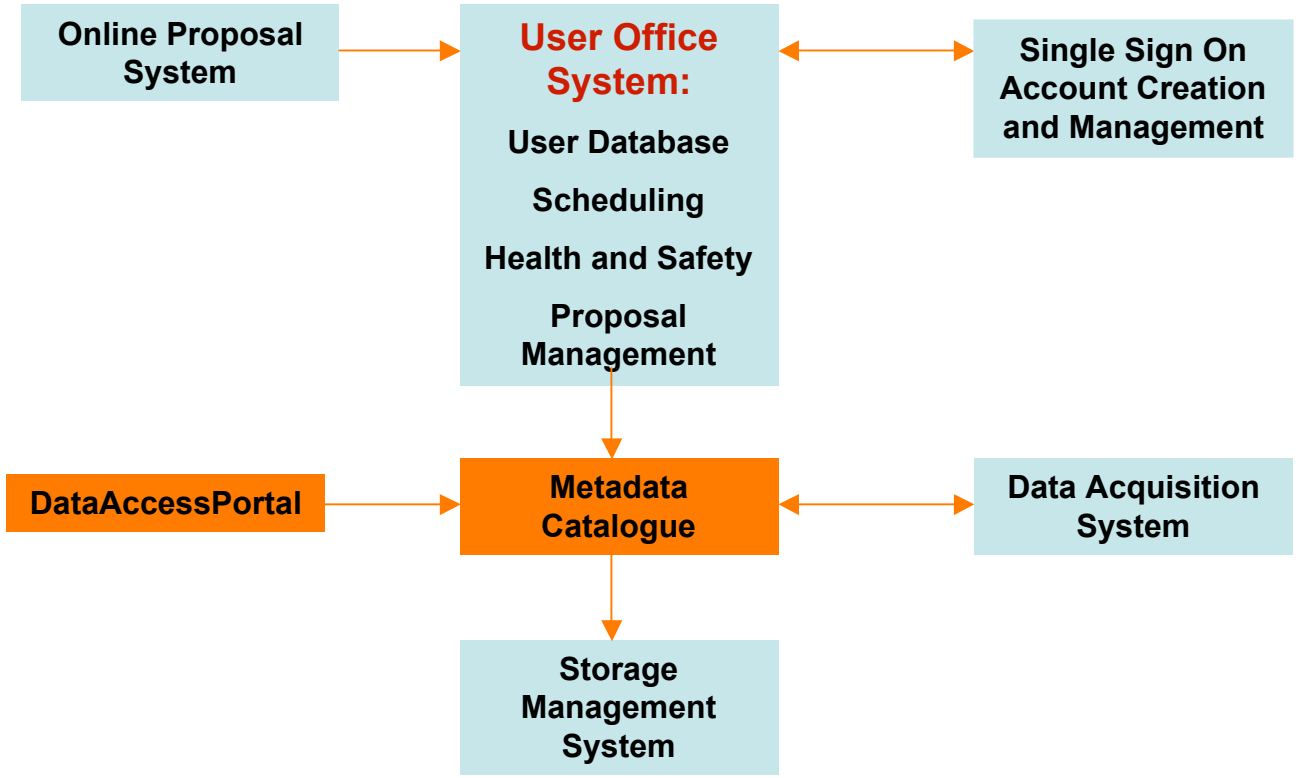
ICAT Principles

The ICAT software suite

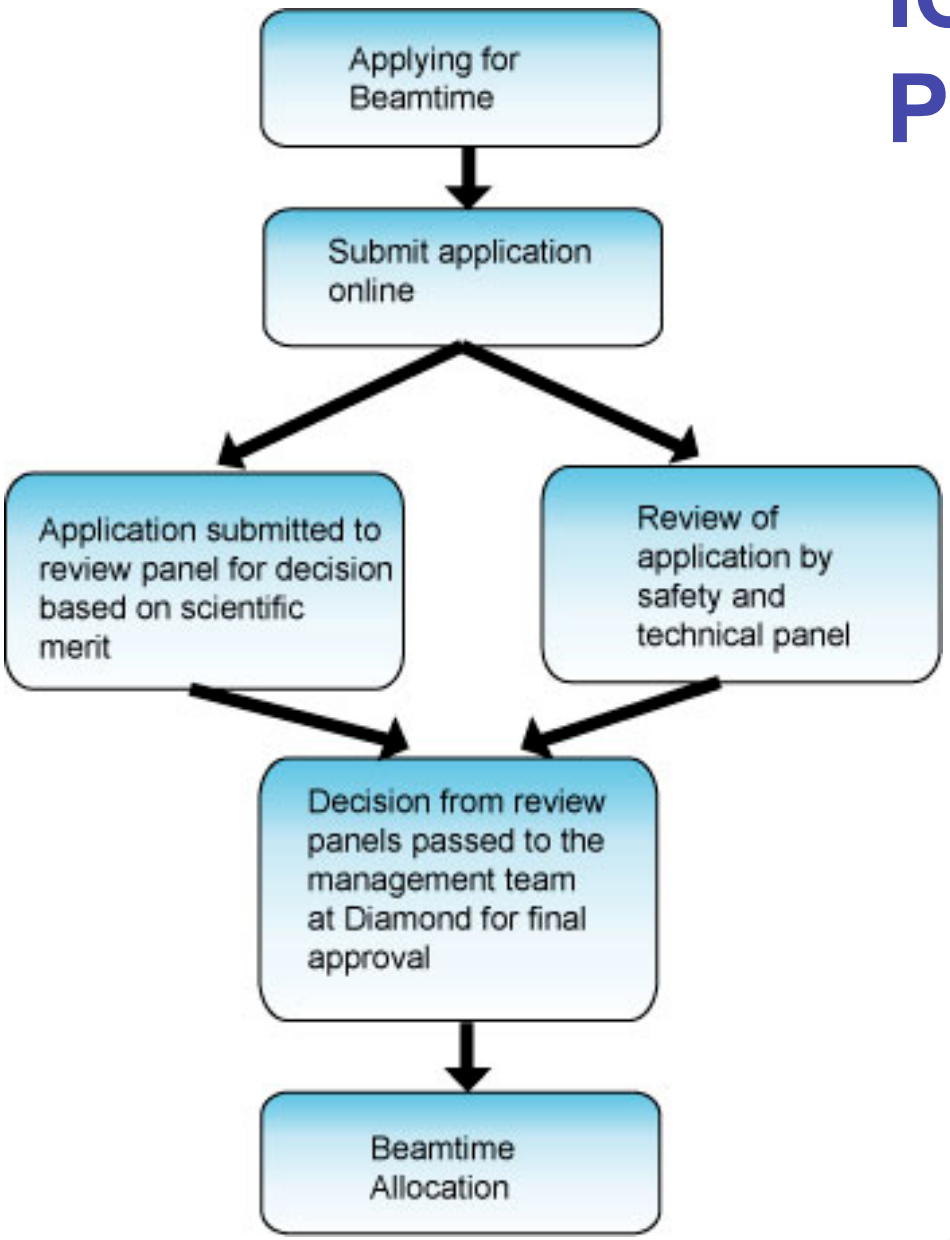
- Catalogues all experiment related information
- Metadata gathered via integration with existing IT systems
  - proposal systems
  - data acquisition
- Provides a well defined API for easy embedding into any applications.



Underlying Data Infrastructure



ICAT Software Suite, providing the crucial integration of key functions.



ICAT and the STFC Proposal Systems

- The entrance point to the Data Management System
- Managed by Facility User Office
- A rich source of contextual information about the users experiment.

ICAT and STFC Data Acquisition

Plug-ins for the data acquisition system ensure automatic, quality controlled collection of data and metadata. ICAT can be easily linked to any existing system.

- **ISIS :**
  - SECI (C#, .net) with link to LabView and openGenie
- **DLS :**
  - Generic Data Acquisition (Java, on top of EPICS)
- **CLF :**
  - For Laser Diagnostics, (LabView)

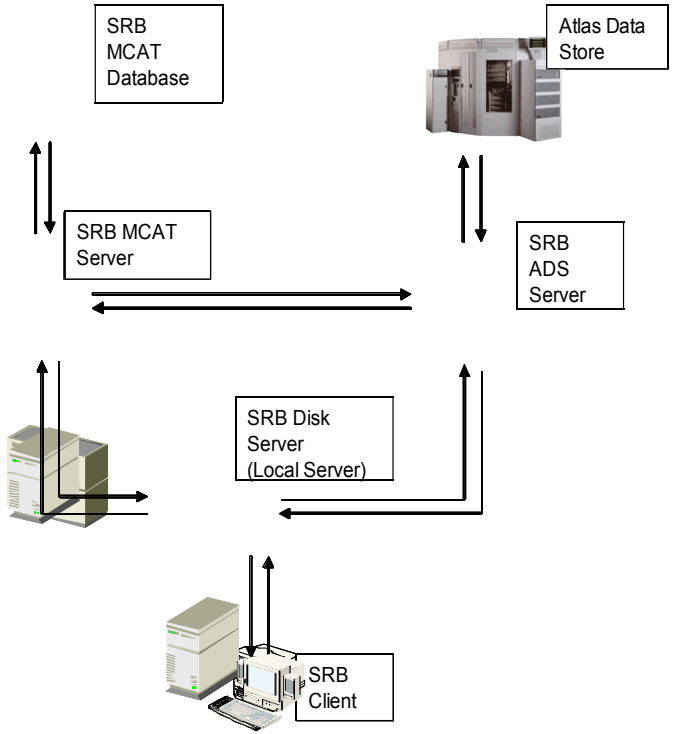
ICAT and Storage Management

Can use the Storage Resource Broker for its Storage Management.

Integrated with ICAT for data access and delivery.

Main advantages :

- **Decoupling physical file location from the logical one.**
- **Strict Security**
- **Expandable to many storage systems**



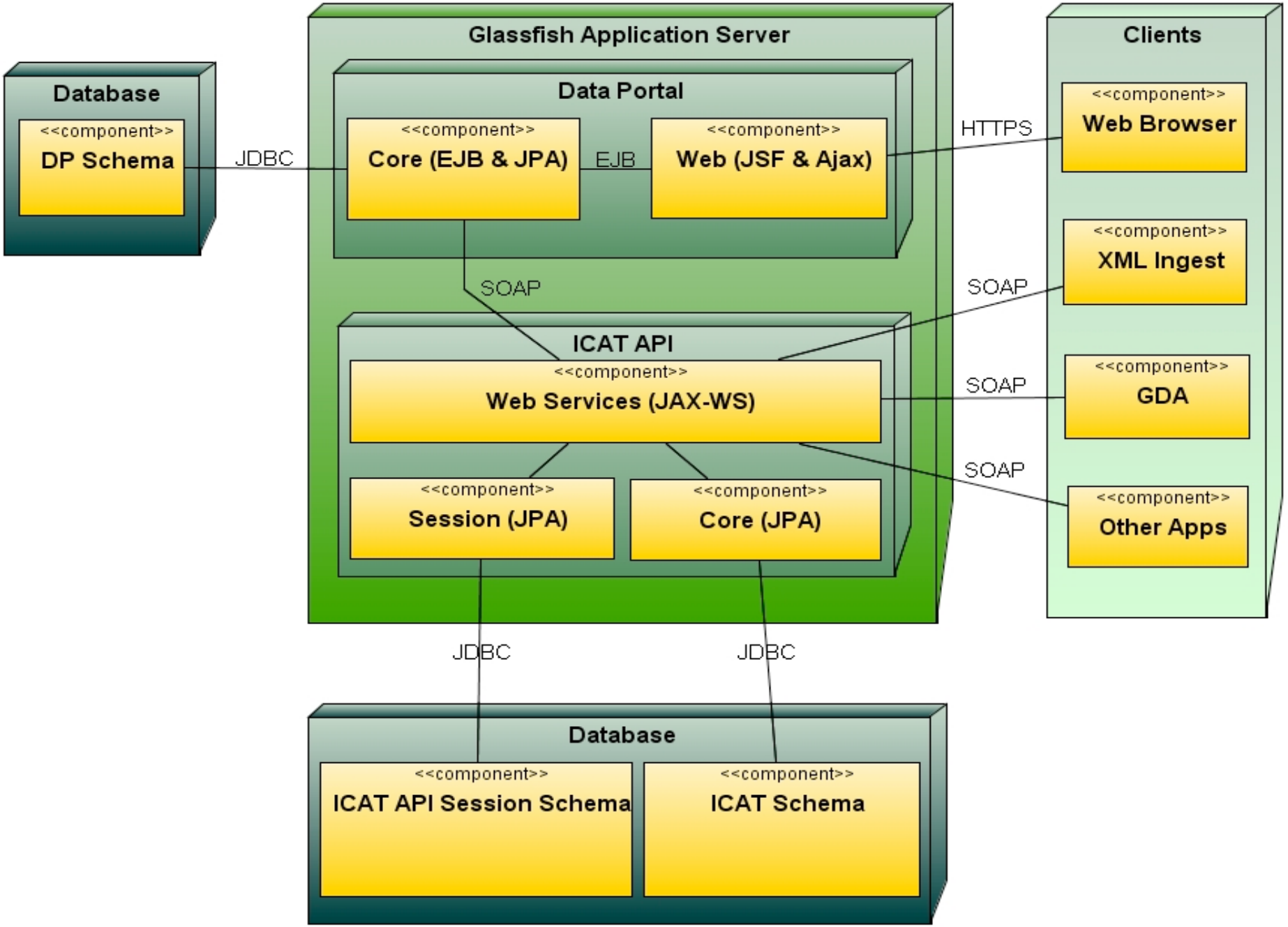
ICAT 3.3 Components

ICAT API Version 3.3

- **the Grid aware software infrastructure**
- **enables applications to interface to the ICAT catalogue.**

Data Portal Version 3.3

- **Data Search and Retrieval (DSR) requirements of the STFC.**
- **Uses ICAT API 3.3 .**


Glassfish Application Server
Database
<<component>>
DP Schema
JDBC
Data Portal
<<component>>
Core (EJB & JPA)
EJB
<<component>>
Web (JSF & Ajax)
HTTPS
Clients
<<component>>
Web Browser
SOAP
ICAT API
<<component>>
Web Services (JAX-WS)
SOAP
<<component>>
XML Ingest
SOAP
<<component>>
GDA
SOAP
<<component>>
Other Apps
<<component>>
Session (JPA)
<<component>>
Core (JPA)
JDBC
JDBC
Database
<<component>>
ICAT API Session Schema
<<component>>
ICAT Schema

DataPortal Walkthrough

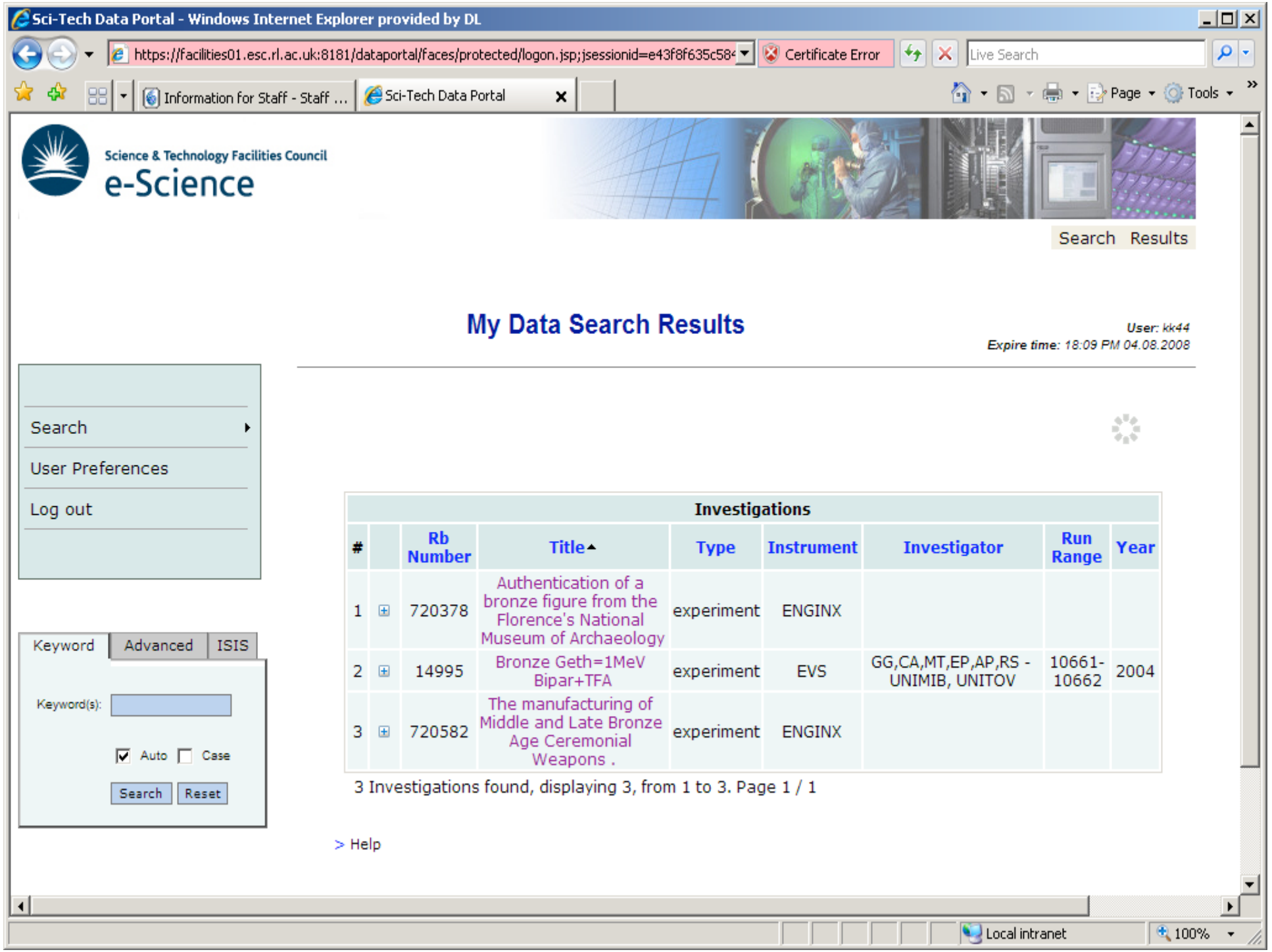
DataPortal for ICAT Version 3.3

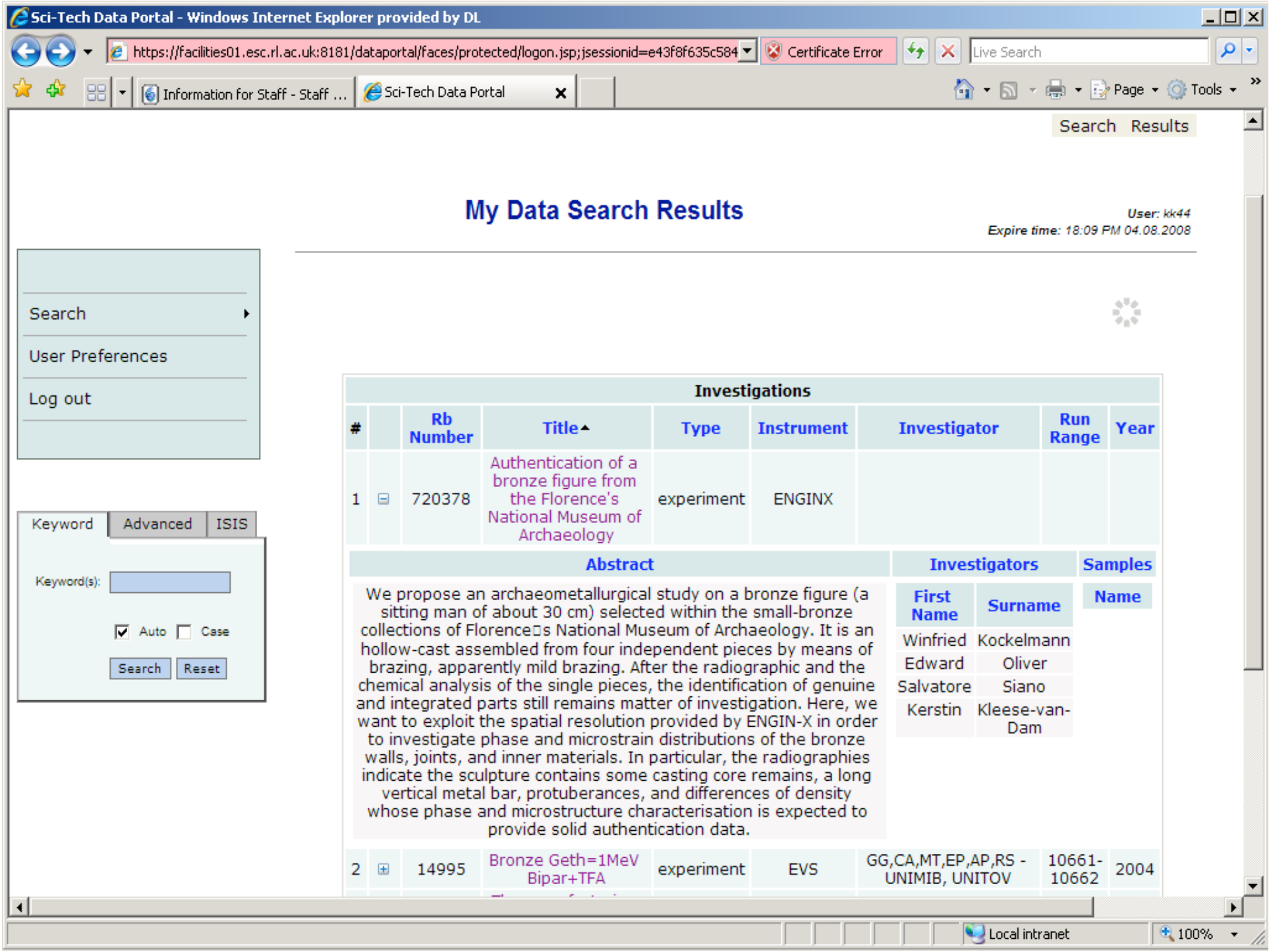
The DataPortal is a customisable web interface to interact with the ICAT version 3.3.

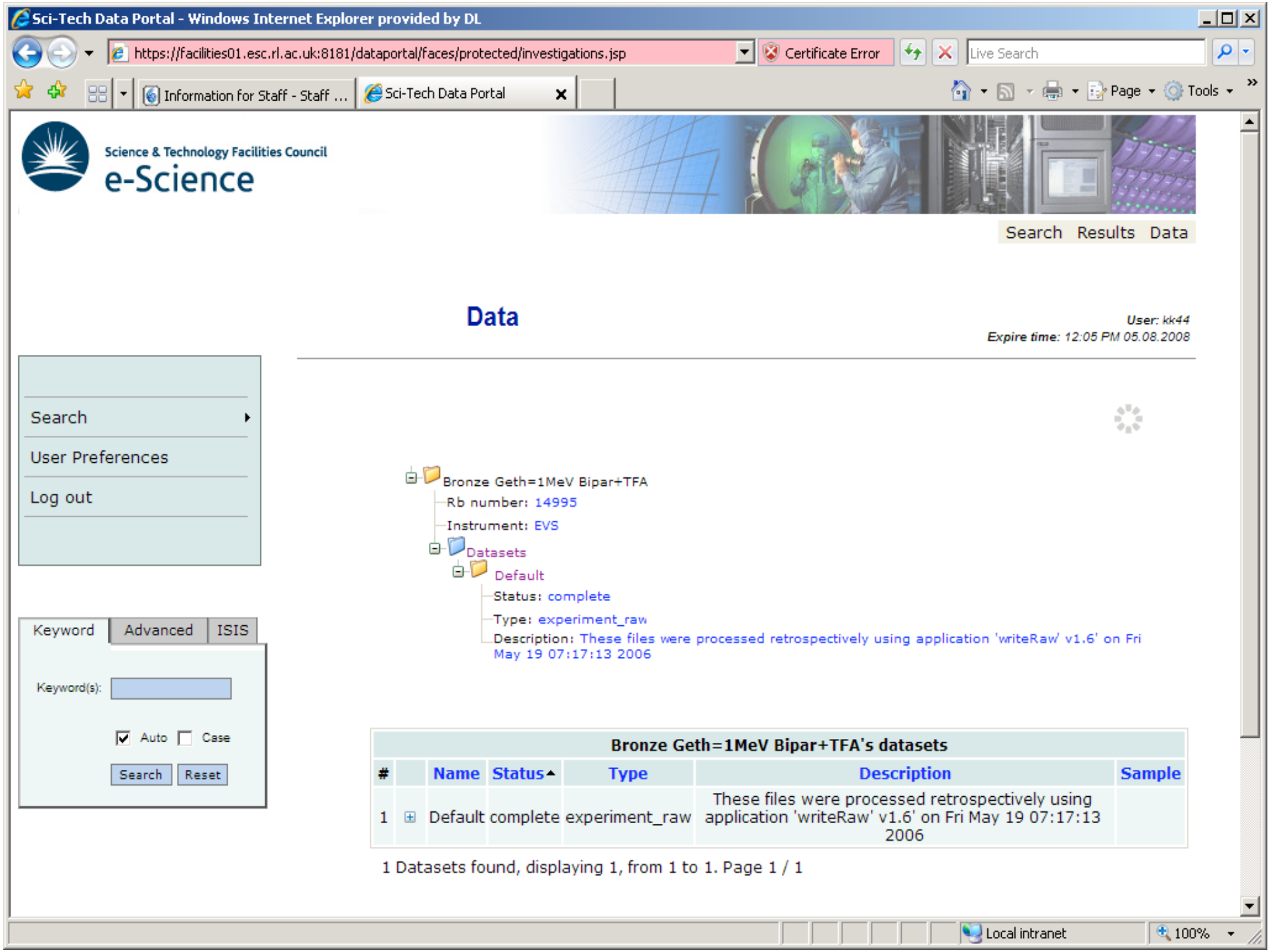
There are at present two distinctive versions one for ISIS and one for DLS.

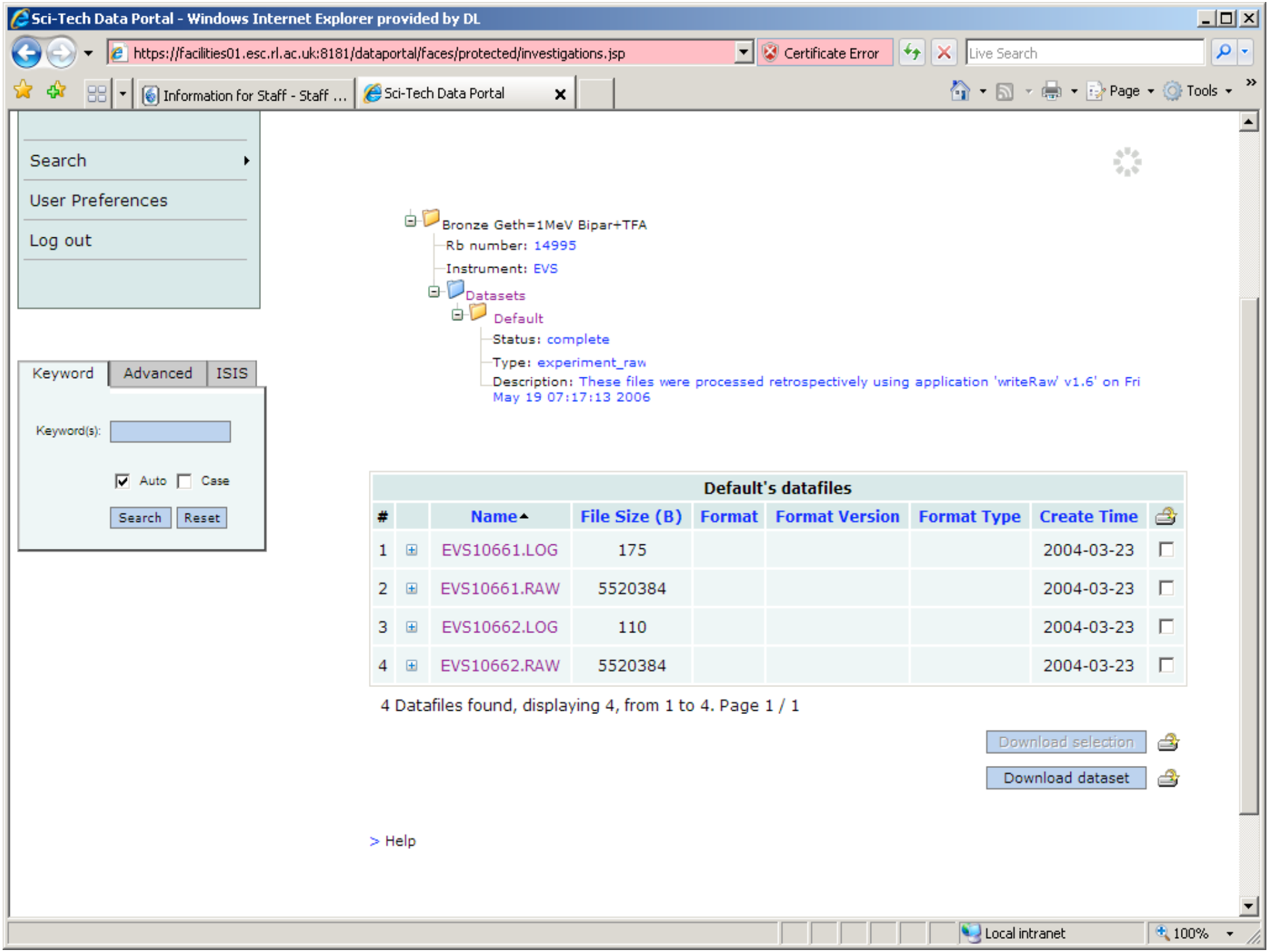
- **the underlying functionality is the same**
- **the graphical representation and choice of used services varies.**

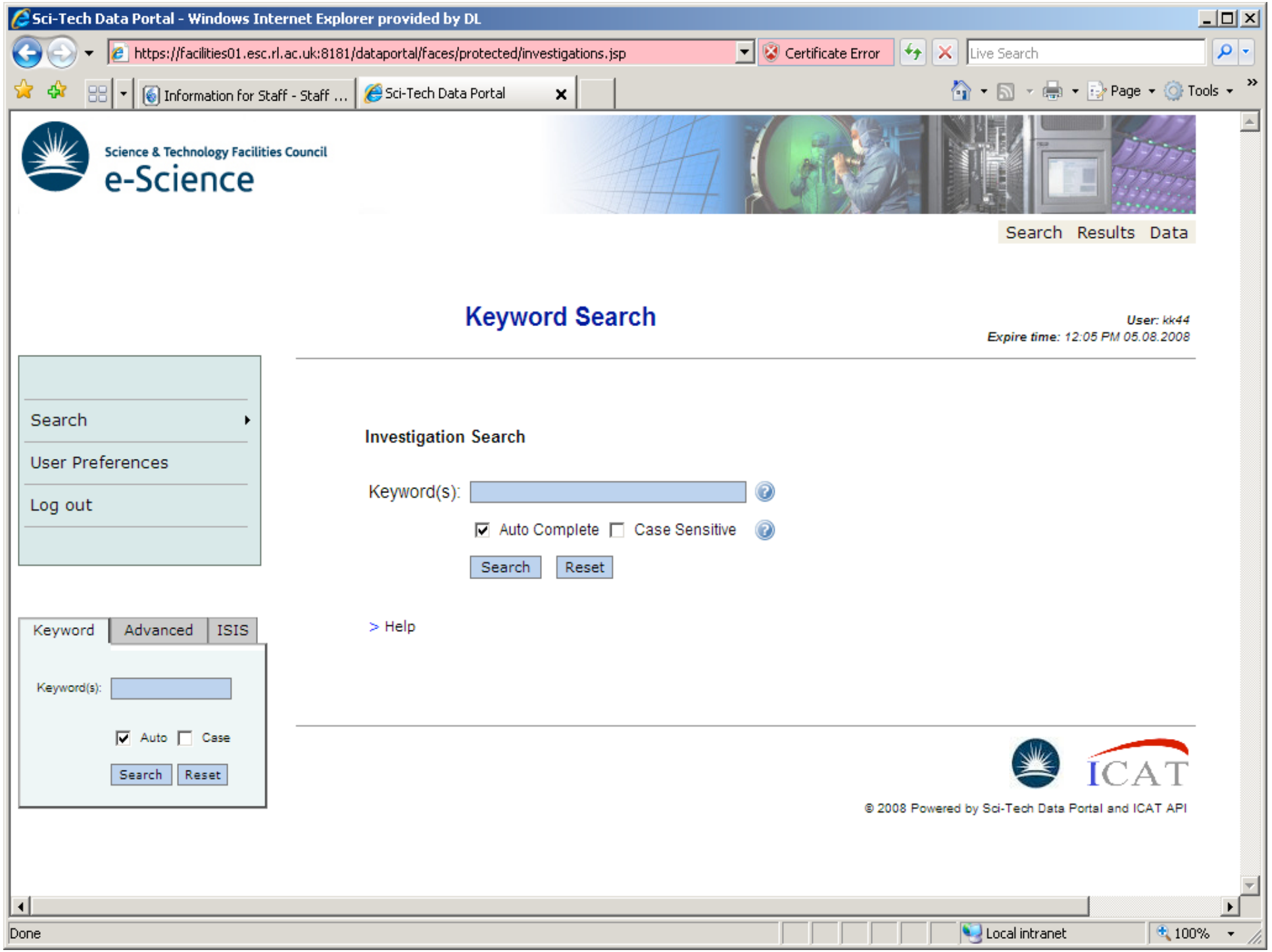
The DataPortal offers a number of search interfaces, the ability to explore investigations and download associated data.

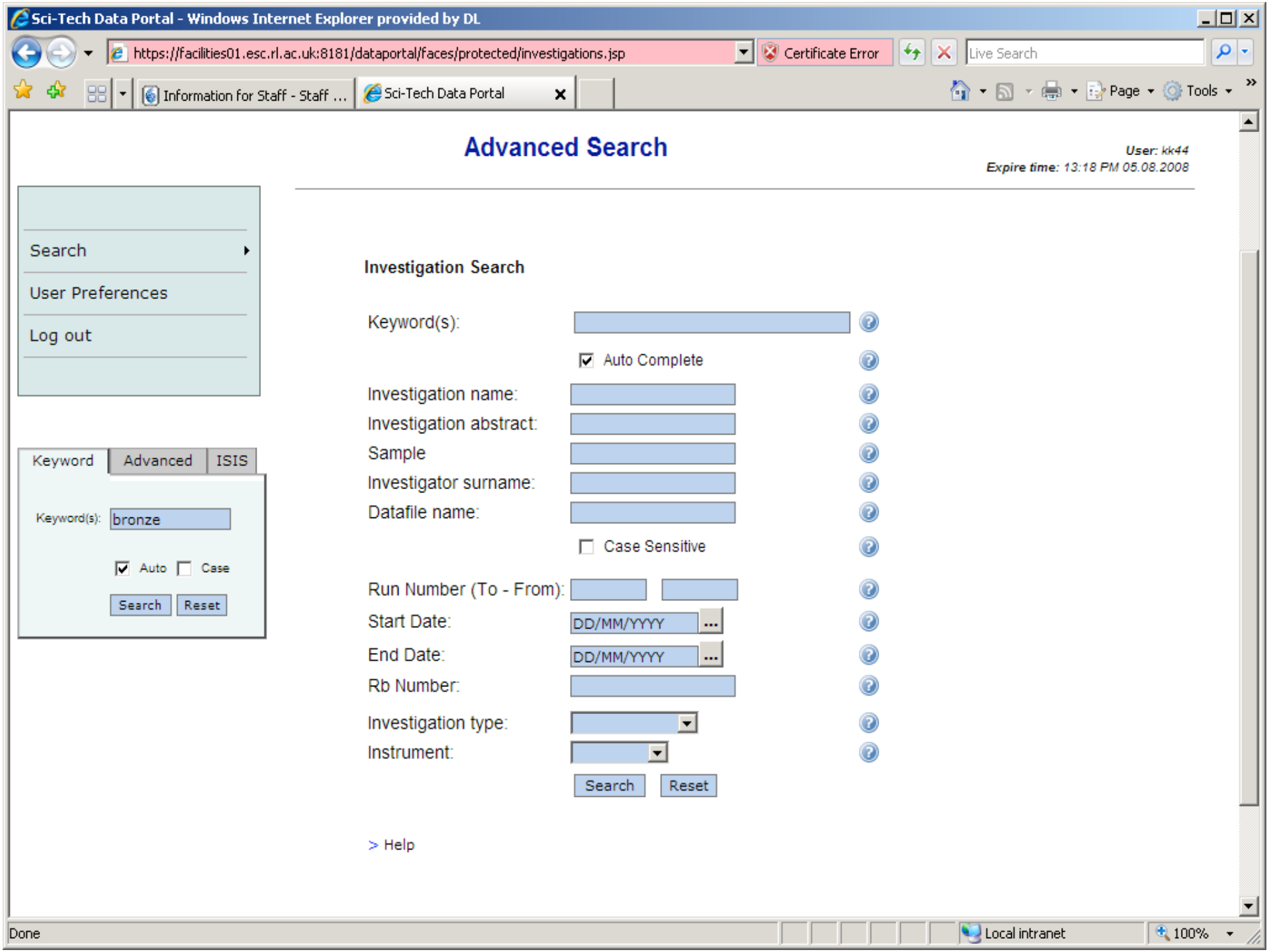

Sci-Tech Data Portal - Windows Internet Explorer provided by DL
https://facilities01.esc.rl.ac.uk:8181/dataportal/faces/protected/logon.jsp;jsessionid=e43f8f635c58
Certificate Error
Live Search
Information for Staff - Staff ...
Sci-Tech Data Portal
Page
Tools
Science & Technology Facilities Council
e-Science
Search Results
My Data Search Results
User: kk44
Expire time: 18:09 PM 04.08.2008
Search
User Preferences
Log out
Keyword
Advanced
ISIS
Keyword(s):
Auto
Case
Search
Reset
Investigations
Rb Number Title Type Instrument Investigator Run Range Year
1 720378 Authentication of a bronze figure from the Florence's National Museum of Archaeology experiment ENGINX
2 14995 Bronze Geth=1MeV Bipar+TFA experiment EVS GG,CA,MT,EP,AP,RS - UNIMIB, UNITOV 10661-10662 2004
3 720582 The manufacturing of Middle and Late Bronze Age Ceremonial Weapons . experiment ENGINX
3 Investigations found, displaying 3, from 1 to 3. Page 1 / 1
> Help
Local intranet
100%

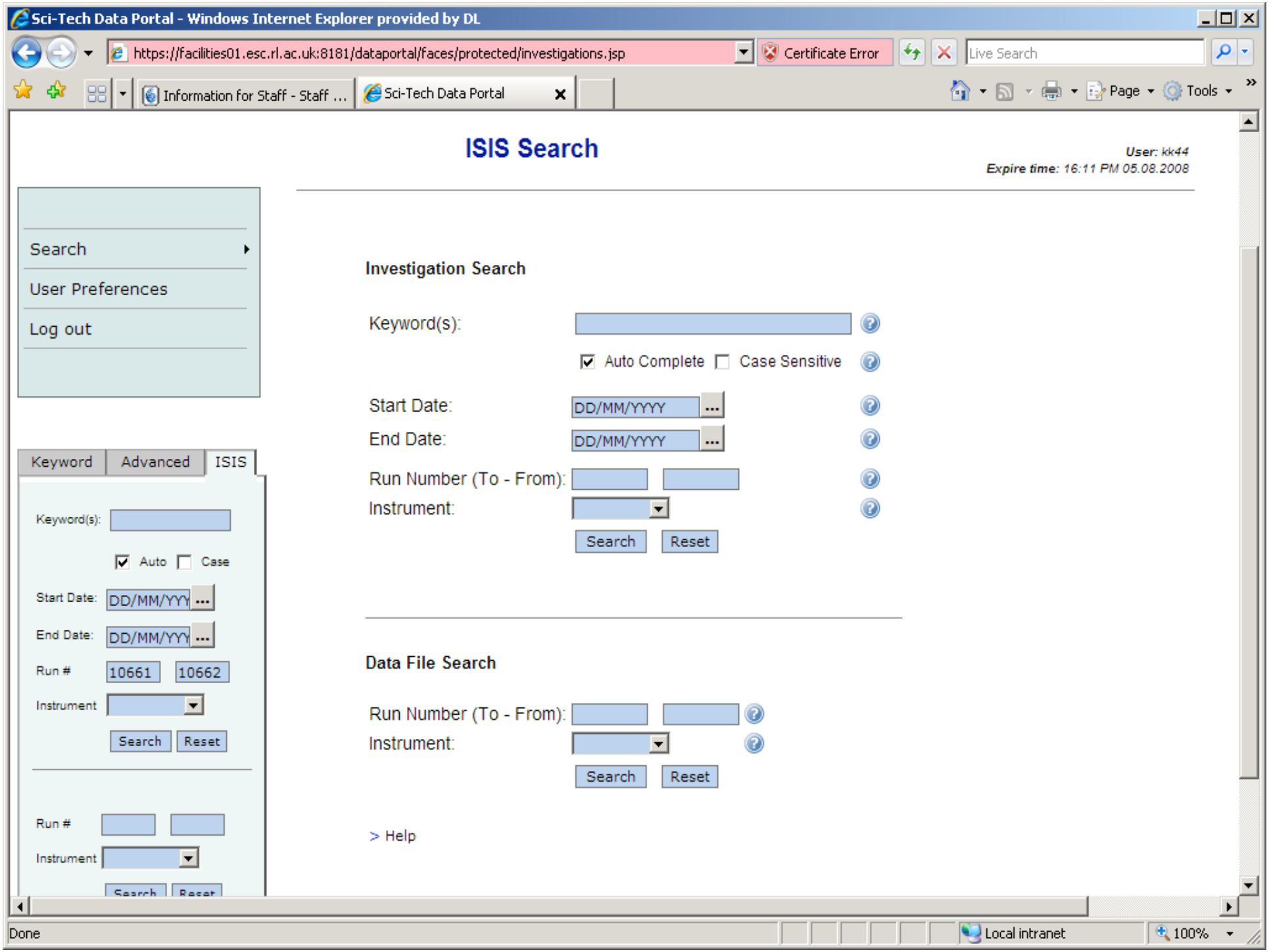

Sci-Tech Data Portal - Windows Internet Explorer provided by DL
https://facilities01.esc.rl.ac.uk:8181/dataportal/faces/protected/logon.jsp;jsessionid=e43f8f635c584
Certificate Error
Live Search
Information for Staff - Staff ...
Sci-Tech Data Portal
Page
Tools
Search Results
My Data Search Results
User: kk44
Expire time: 18:09 PM 04.08.2008
Search
User Preferences
Log out
Keyword
Advanced
ISIS
Keyword(s):
Auto
Case
Search
Reset
Investigations
Rb Number Title Type Instrument Investigator Run Range Year
1 720378 Authentication of a bronze figure from the Florence's National Museum of Archaeology experiment ENGINX
Abstract
We propose an archaeometallurgical study on a bronze figure (a sitting man of about 30 cm) selected within the small-bronze collections of Florence□s National Museum of Archaeology. It is an hollow-cast assembled from four independent pieces by means of brazing, apparently mild brazing. After the radiographic and the chemical analysis of the single pieces, the identification of genuine and integrated parts still remains matter of investigation. Here, we want to exploit the spatial resolution provided by ENGIN-X in order to investigate phase and microstrain distributions of the bronze walls, joints, and inner materials. In particular, the radiographies indicate the sculpture contains some casting core remains, a long vertical metal bar, protuberances, and differences of density whose phase and microstructure characterisation is expected to provide solid authentication data.
Investigators
First Name Surname
Winfried Kockelmann
Edward Oliver
Salvatore Siano
Kerstin Kleese-van-Dam
Samples
Name
2 14995 Bronze Geth=1MeV Bipar+TFA experiment EVS GG,CA,MT,EP,AP,RS - UNIMIB, UNITOV 10661-10662 2004
Local intranet
100%


Sci-Tech Data Portal - Windows Internet Explorer provided by DL
https://facilities01.esc.rl.ac.uk:8181/dataportal/faces/protected/investigations.jsp
Certificate Error
Live Search
Information for Staff - Staff ...
Sci-Tech Data Portal
Page
Tools
Science & Technology Facilities Council
e-Science
Search Results Data
Data
User: kk44
Expire time: 12:05 PM 05.08.2008
Search
User Preferences
Log out
Keyword
Advanced
ISIS
Keyword(s):
Auto
Case
Search
Reset
Bronze Geth=1MeV Bipar+TFA
Rb number: 14995
Instrument: EVS
Datasets
Default
Status: complete
Type: experiment_raw
Description: These files were processed retrospectively using application 'writeRaw' v1.6' on Fri May 19 07:17:13 2006
Bronze Geth=1MeV Bipar+TFA's datasets
Name Status Type Description Sample
1 Default complete experiment_raw These files were processed retrospectively using application 'writeRaw' v1.6' on Fri May 19 07:17:13 2006
1 Datasets found, displaying 1, from 1 to 1. Page 1 / 1
Local intranet
100%


Sci-Tech Data Portal - Windows Internet Explorer provided by DL
https://facilities01.esc.rl.ac.uk:8181/dataportal/faces/protected/investigations.jsp
Certificate Error
Live Search
Information for Staff - Staff ...
Sci-Tech Data Portal
Page
Tools
Search
User Preferences
Log out
Keyword
Advanced
ISIS
Keyword(s):
Auto
Case
Search
Reset
Bronze Geth=1MeV Bipar+TFA
Rb number: 14995
Instrument: EVS
Datasets
Default
Status: complete
Type: experiment_raw
Description: These files were processed retrospectively using application 'writeRaw' v1.6' on Fri May 19 07:17:13 2006
Default's datafiles
Name File Size (B) Format Format Version Format Type Create Time
1 EVS10661.LOG 175 2004-03-23
2 EVS10661.RAW 5520384 2004-03-23
3 EVS10662.LOG 110 2004-03-23
4 EVS10662.RAW 5520384 2004-03-23
4 Datafiles found, displaying 4, from 1 to 4. Page 1 / 1
Download selection
Download dataset
> Help
Local intranet
100%


Sci-Tech Data Portal - Windows Internet Explorer provided by DL
https://facilities01.esc.rl.ac.uk:8181/dataportal/faces/protected/investigations.jsp
Certificate Error
Live Search
Information for Staff - Staff ...
Sci-Tech Data Portal
Page
Tools
Science & Technology Facilities Council
e-Science
Search Results Data
Keyword Search
User: kk44
Expire time: 12:05 PM 05.08.2008
Search
User Preferences
Log out
Investigation Search
Keyword(s):
Auto Complete
Case Sensitive
Search
Reset
> Help
Keyword
Advanced
ISIS
Keyword(s):
Auto
Case
Search
Reset
ICAT
© 2008 Powered by Sci-Tech Data Portal and ICAT API
Done
Local intranet
100%


Sci-Tech Data Portal - Windows Internet Explorer provided by DL
https://facilities01.esc.rl.ac.uk:8181/dataportal/faces/protected/investigations.jsp
Certificate Error
Live Search
Information for Staff - Staff ...
Sci-Tech Data Portal
Page
Tools
Advanced Search
User: kk44
Expire time: 13:18 PM 05.08.2008
Search
User Preferences
Log out
Keyword
Advanced
ISIS
Keyword(s): bronze
Auto
Case
Search
Reset
Investigation Search
Keyword(s):
Auto Complete
Investigation name:
Investigation abstract:
Sample
Investigator surname:
Datafile name:
Case Sensitive
Run Number (To - From):
Start Date:
DD/MM/YYYY
End Date:
DD/MM/YYYY
Rb Number:
Investigation type:
Instrument:
Search
Reset
> Help
Done
Local intranet
100%


Sci-Tech Data Portal - Windows Internet Explorer provided by DL
https://facilities01.esc.rl.ac.uk:8181/dataportal/faces/protected/investigations.jsp
Certificate Error
Live Search
Information for Staff - Staff ...
Sci-Tech Data Portal
Page
Tools
ISIS Search
User: kk44
Expire time: 16:11 PM 05.08.2008
Search
User Preferences
Log out
Keyword
Advanced
ISIS
Keyword(s):
Auto
Case
Start Date:
DD/MM/YYY
End Date:
DD/MM/YYY
Run #
10661
10662
Instrument
Search
Reset
Run #
Instrument
Search
Reset
Investigation Search
Keyword(s):
Auto Complete
Case Sensitive
Start Date:
DD/MM/YYYY
End Date:
DD/MM/YYYY
Run Number (To - From):
Instrument:
Search
Reset
Data File Search
Run Number (To - From):
Instrument:
Search
Reset
> Help
Done
Local intranet
100%

ISIS Facilities Ontology

Using Ontologies in Metadata

·Ontologies are used to capture knowledge about a domain of interest.

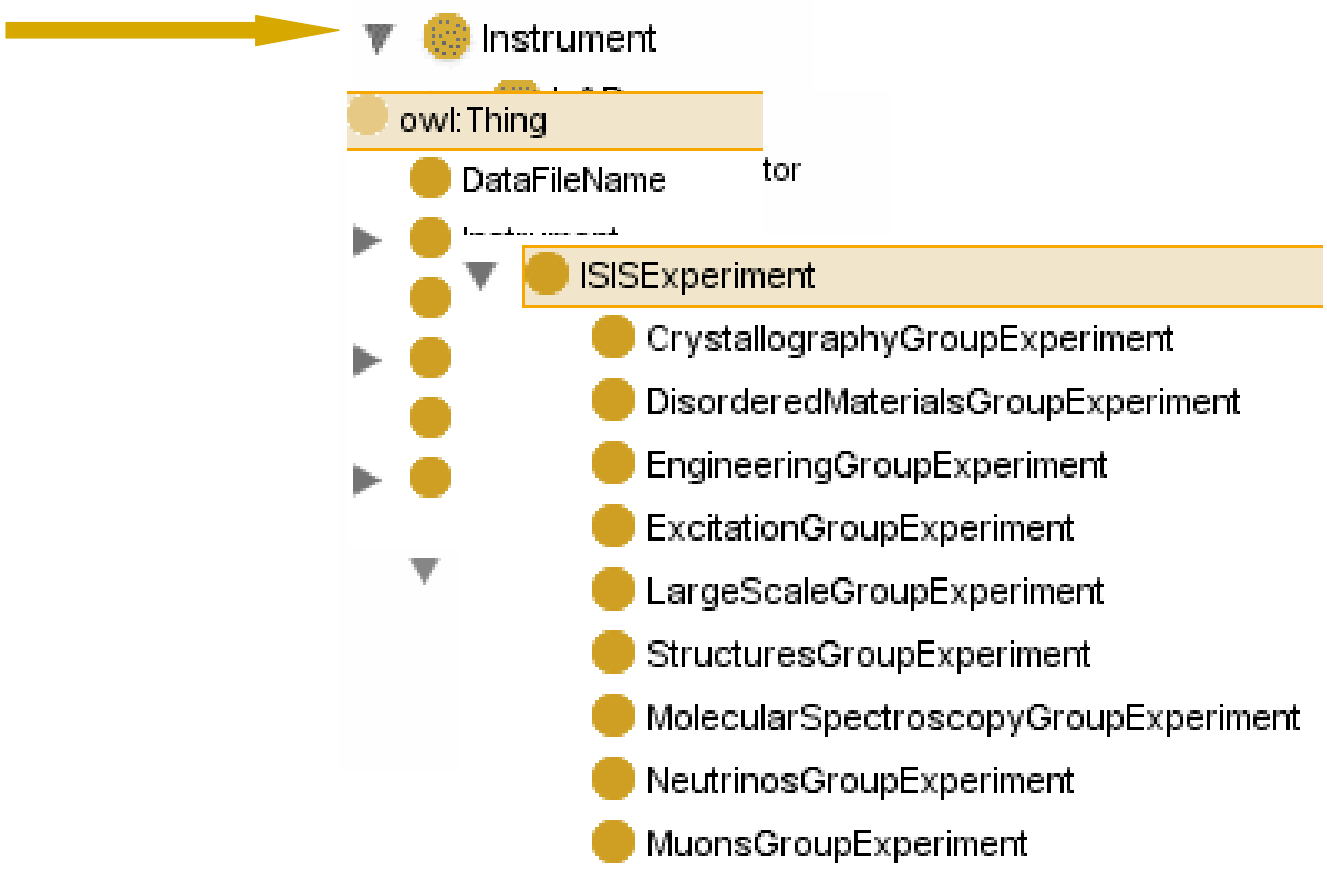
· An ontology describes the concepts in the domain and the relationships that hold between those concepts.

- Provide increased flexibility when representing frequently changing viewpoints of information.
- Alterations can be simply followed up in the model without having to alter the applications on which they are based.
- Allows a unified view of heterogeneous data sources.
- Remove conflicts and terminological uncertainties.
- Facilitate moderated searches, optimisation of the search results.

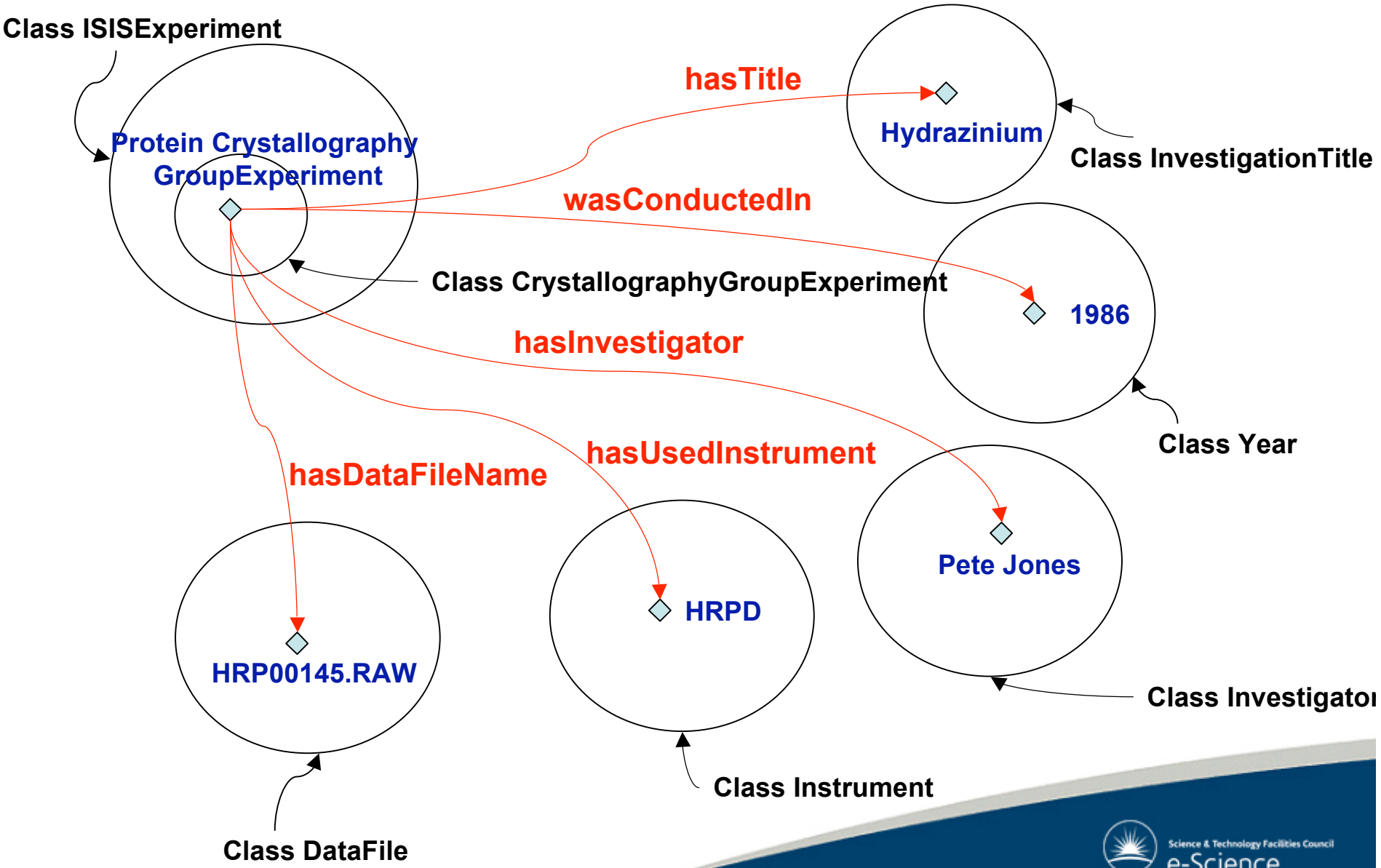
Towards an ISIS Ontology

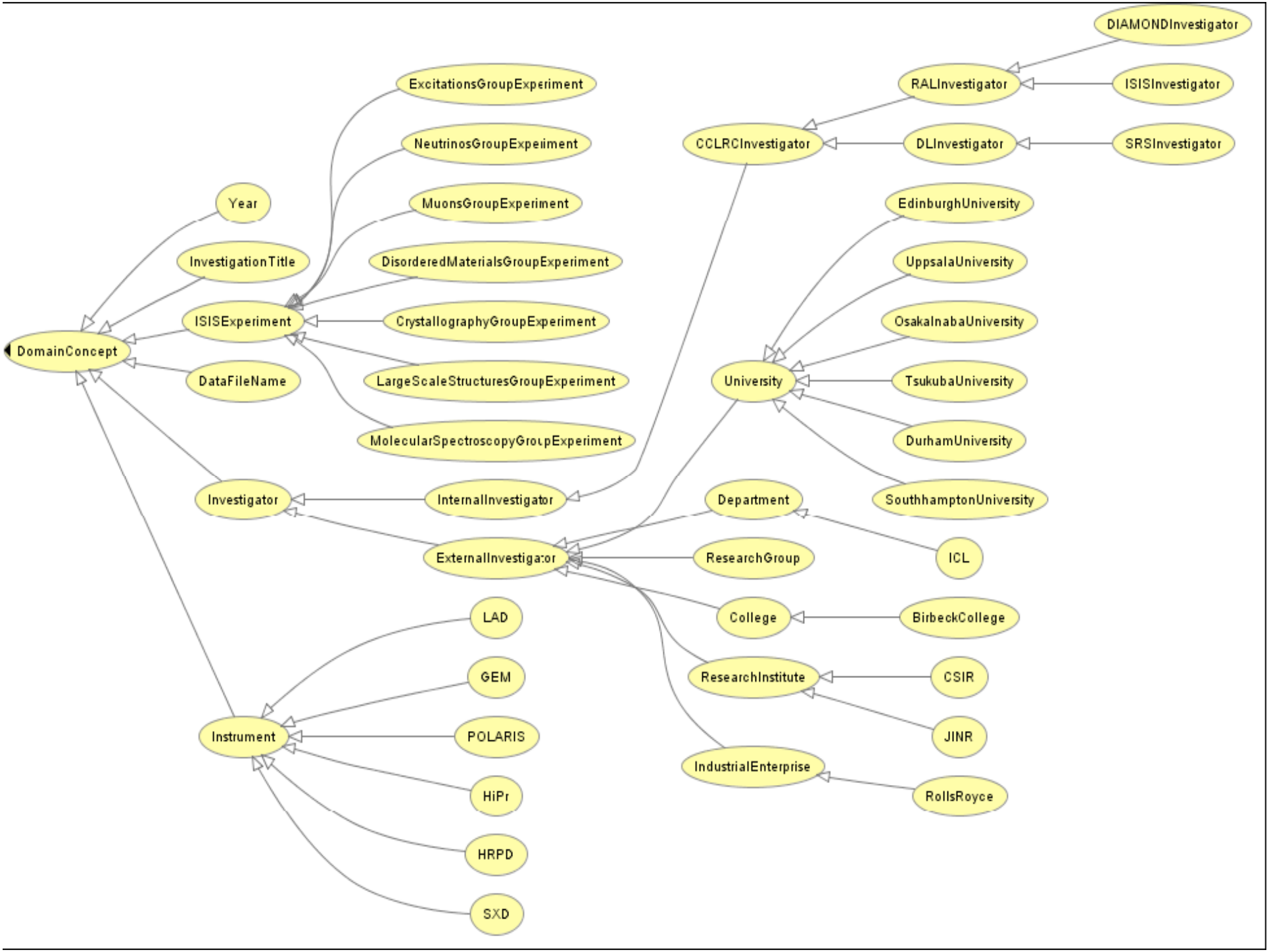
- At present over 10,000 keywords describing experiments are housed in ISIS ICAT
  - many of which are synonyms.
- These keywords are used to index experimental studies, however:
  - free text keywords have no context,
  - hard to map by non-experts to terms used by facilities in the same domain and harder still to those outside.
- The creation of ontologies at ISIS will aid in the mapping of concrete manifestations of familiar terms in one domain as well as related concepts in different domains.
- This will facilitate searching of data by category and grouping of data into keywords across studies.
- This could aid in the cross facility searching of related scientific data from the various scientific facilities housed at STFC e.g. CLF and DLS
- Work of Louisa Casely-Hayford

ISIS Facilities Ontology Hierarchy



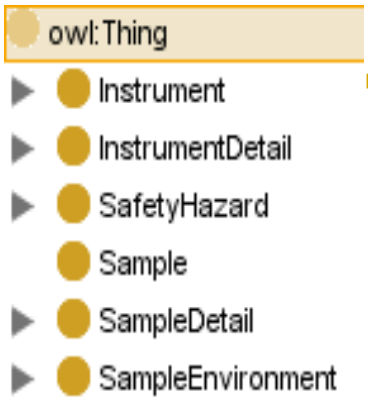
ISIS Facilities Ontology



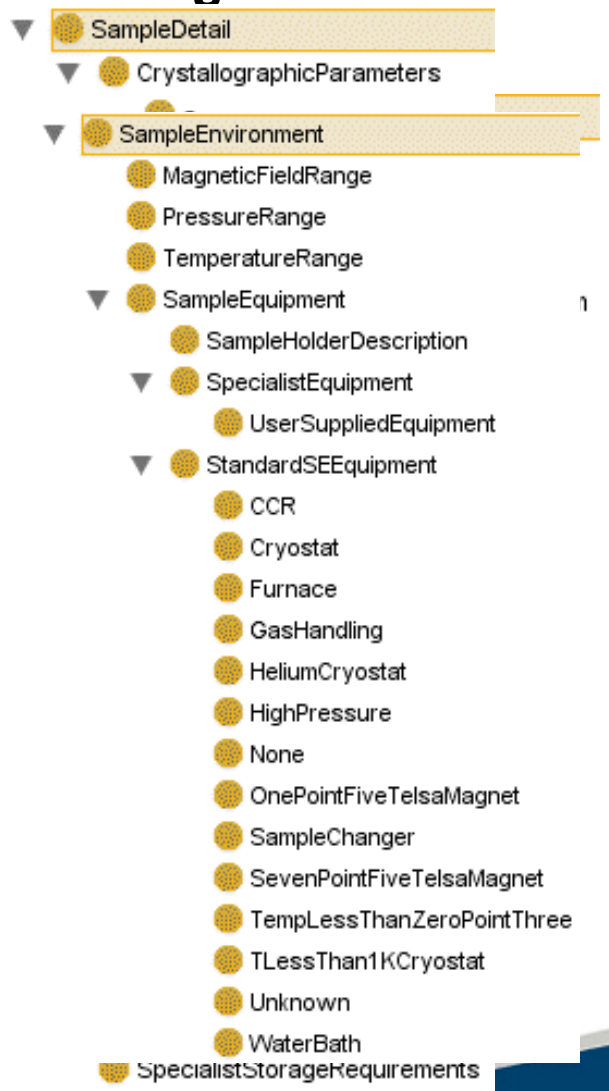

DIAMONDInvestigator
RALInvestigator
ISISInvestigator
ExcitationsGroupExperiment
CCLRCInvestigator
DLInvestigator
SRSInvestigator
NeutrinosGroupExperiment
Year
MuonsGroupExperiment
EdinburghUniversity
InvestigationTitle
DisorderedMaterialsGroupExperiment
UppsalaUniversity
ISISExperiment
CrystallographyGroupExperiment
OsakaInabaUniversity
DomainConcept
University
DataFileName
LargeScaleStructuresGroupExperiment
TsukubaUniversity
DurhamUniversity
MolecularSpectroscopyGroupExperiment
SouthhamptonUniversity
Investigator
InternalInvestigator
Department
ExternalInvestigator
ResearchGroup
ICL
College
BirbeckCollege
LAD
GEM
ResearchInstitute
CSIR
Instrument
POLARIS
JINR
IndustrialEnterprise
HiPr
RollsRoyce
HRPD
SXD

Sample, Investigator and Experiment Ontologies

Sample



Investigator



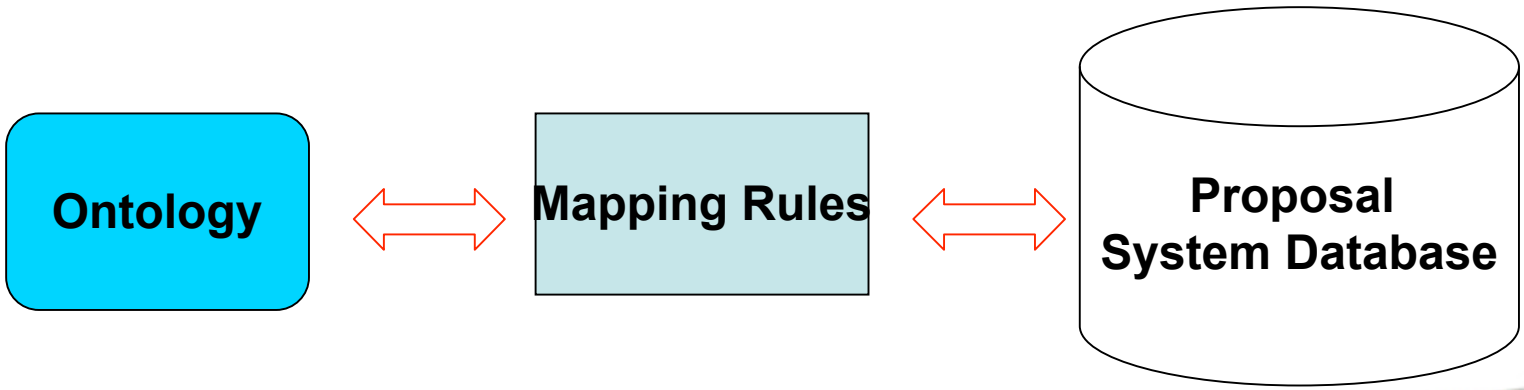
Experiment

Science & Technology Facilities Council
e-Science

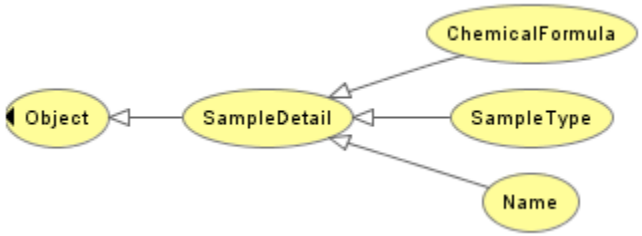
Topic Mapping Tool

- Mapping Tool provides a way of linking proposal system data to the structure of the ontology.

- Data is mapped to the ontology structure according to a set of defined rules.

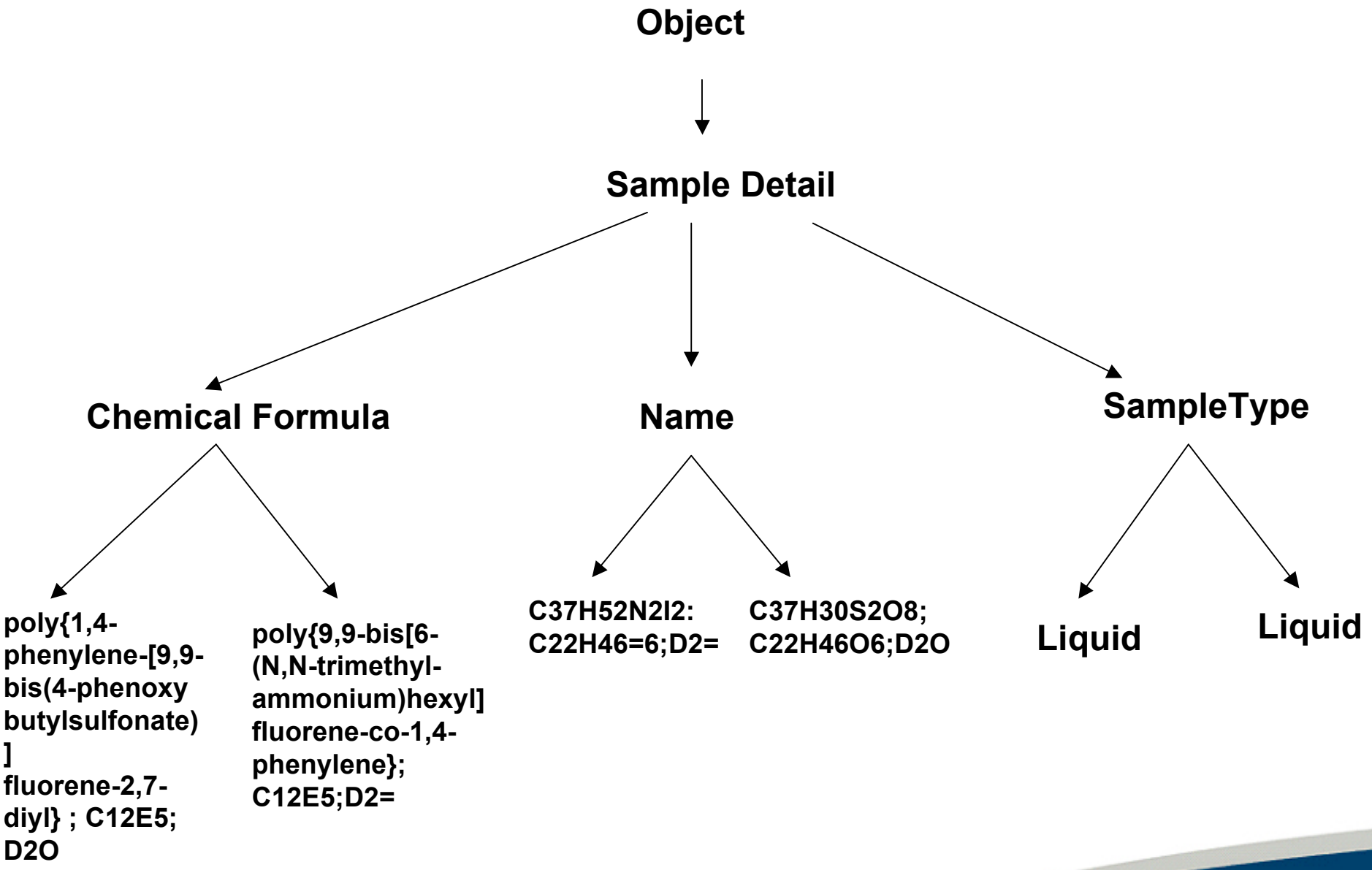


Mapping Tool



ID	NAME	PARENT_ID
2	ChemicalFormula	1
3	Name	1
4	SampleType	1
5	C37H30S2O8; C22H46O6;D2O	2
6	C37H52N2I2:C22H46=6;D2=	2
7	poly{1,4-phenylene-[9,9-bis(4-phenoxy-butylsulfonate)]fluorene-2,7-diyl} ; C12E5; D2O	3
8	poly{9,9-bis[6-(N,N-trimethyl-ammonium)hexyl]fluorene-co-1,4-phenylene};C12E5;D2=	3
9	Liquid	4
10	Liquid	4
0	Object	-1
1	SampleDetail	0

STUDY_ID	TOPIC_ID
639	7
639	8
639	9
639	10
639	5
639	6


Object
Sample Detail
Chemical Formula
Name
SampleType
poly{1,4-phenylene-[9,9-bis(4-phenoxy butylsulfonate)] fluorene-2,7-diyl} ; C12E5; D2O
poly{9,9-bis[6-(N,N-trimethyl-ammonium)hexyl] fluorene-co-1,4-phenylene}; C12E5;D2=
C37H52N2I2: C22H46=6;D2=
C37H30S2O8; C22H46O6;D2O
Liquid
Liquid

CSMD Current and Future Direction

What we have achieved

Agreed common metadata and data formats and single sign on allow scientist to have rapid access at any stage of their work

- **A 20 year back catalogue of ISIS raw data is available.**

- **All future data collected at STFC Facilities and DLS will be curated and made available for reuse now and in the future.**

Creating a powerful, long lasting scientific knowledge resource.

ICAT Usage in ISIS

- 22 neutron and muon instruments are
  - populating ICAT in real time at an average of 330 datafiles per hour.
  - 3,133,639 files (as of 9 Oct 08) that are indexed by the ISIS ICAT
  - ~4 Tb in terms of data volume..
- The new Target Station 2 at ISIS be entered into ICAT in exactly the same way as TS1.
- There are in the region of 800 experiments/investigations performed at ISIS each year.

Core Scientific Metadata Model Usage

Used on many projects

- STFC DataPortal
  - XML Schema Implementation
  - Serving data from
    - MPIM (Max-Planck-Institut für Meteorologie, Hamburg)
    - STFC Facilities (ISIS, DLS, BADC-test)
- NERC funded 'Environment from the Molecular'
  - Mini-Grid DataPortal Transport Layer
- EPSRC funded 'Simulation of complex materials'
  - Mini-Grid DataPortal Transport Layer

Other projects and CSMD

EPSRC funded MyGrid BioInformatics project

- information model based on version 1 of the CSMD Model

This is being taken in the *myIB project*

- Application to Integrated biology
- Extending to provenance tracking in computational steering

Also influence projects:

- Comb-eChem, eBank, eCrystals
- NERC DataGrid

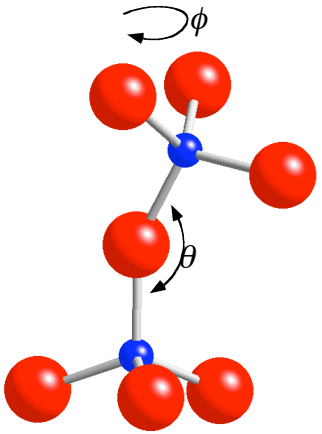
EPSRC CCP1(Collaborative Computational Project in Quantum Chemistry)

- assessing CSMD for metadata needs on their Grid Data Management Middleware

CSMD and Simulations

Detailed Information about the Data Analysis and underpinning Computational Simulations:

- **Simulation/Analysis Code and version**
- **Simulation Set-up and Parameters**
- **Information about the Compute Resource**
- **Key Parameter from the Simulation Results**
- **Keywords and Classifications**



Sharing Publications

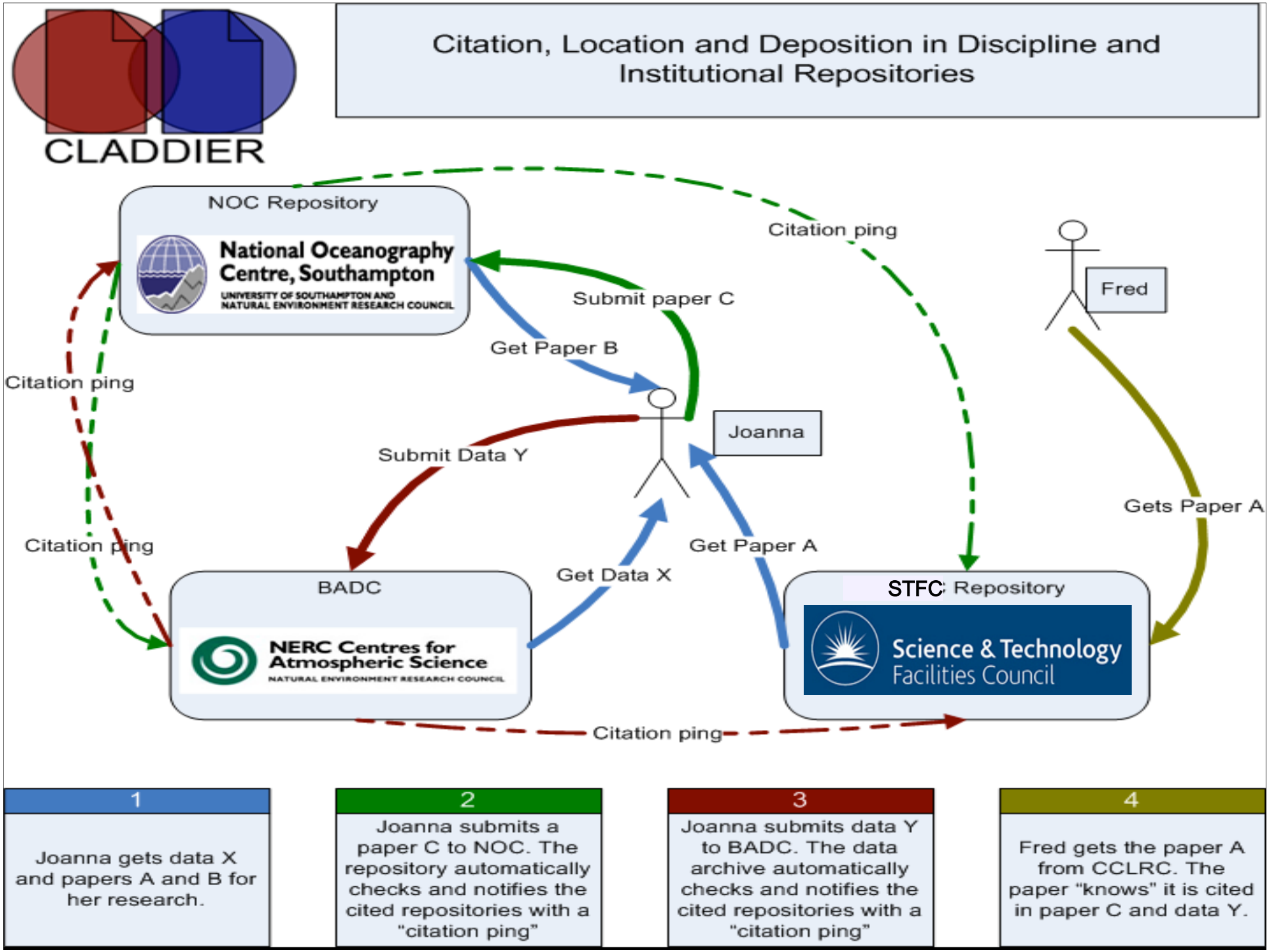
CSMD offers the potential to integrate the outputs of scientific research: data and publications.

Institutional Repository s/w now very well established

- ePrints, DSpace, Fedora, ePubs
- Large body of expertise available
- Standard metadata models and protocols:
  - DC-APs, FRBR, OAI-PMH, OAI-ORE
- Not yet embedded in science practise
  - except HEP!

Linking science data and publications

- Not yet well established
- Needs data citation
- Needs peer review of data
- Can (and should) be done on a P2P basis


CLADDIER
Citation, Location and Deposition in Discipline and Institutional Repositories
NOC Repository
National Oceanography Centre, Southampton
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL
Citation ping
Submit paper C
Get Paper B
Fred
Joanna
Submit Data Y
Gets Paper A
Citation ping
Get Paper A
Get Data X
BADC
NERC Centres for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL
STFC Repository
Science & Technology Facilities Council
Citation ping
1
Joanna gets data X and papers A and B for her research.
2
Joanna submits a paper C to NOC. The repository automatically checks and notifies the cited repositories with a "citation ping"
3
Joanna submits data Y to BADC. The data archive automatically checks and notifies the cited repositories with a "citation ping"
4
Fred gets the paper A from CCLRC. The paper "knows" it is cited in paper C and data Y.

Mapping between Dublin Core and CSMD

Title
- Study: Name

Creator
- Study: Investigator: Name (Role is principle investigator)

Subject
- Topic: Keyword

Description
- Study: Study Information: Purpose

Publisher
- Investigation: Data Manager

Contributor
- Study: Investigator: Name ; Investigation: Data Manager

Date
- Study: Study Information: Time

Resource Type
- *Collection*; or *Dataset.*

Format
- Data Description: File Format

Resource Identifier
- Study: Study Id (whole study)
- Data description: File: URI (for individual data files).

Source
- Data description: Data sets: Related Data sets
- Related Material: Related work

Language
- *Not covered in the current metadata format; but an simple extension*

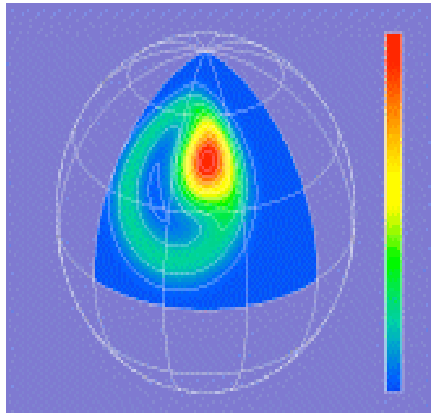
Relation
- Related Material: Related work

Coverage
- Data description: Logical Description: Coverage

Rights Management
- Access Conditions

Towards an Application profile for Science Data

Future Challenges



- **Data Policy and Ownership.**

- **Data and Metadata Curation for long-term reuse.**

- **Enabling scientist to use information from unfamiliar methods.**

- **Integration with other Projects and Facilities around the world.**

Interoperability

Sharing across boundaries

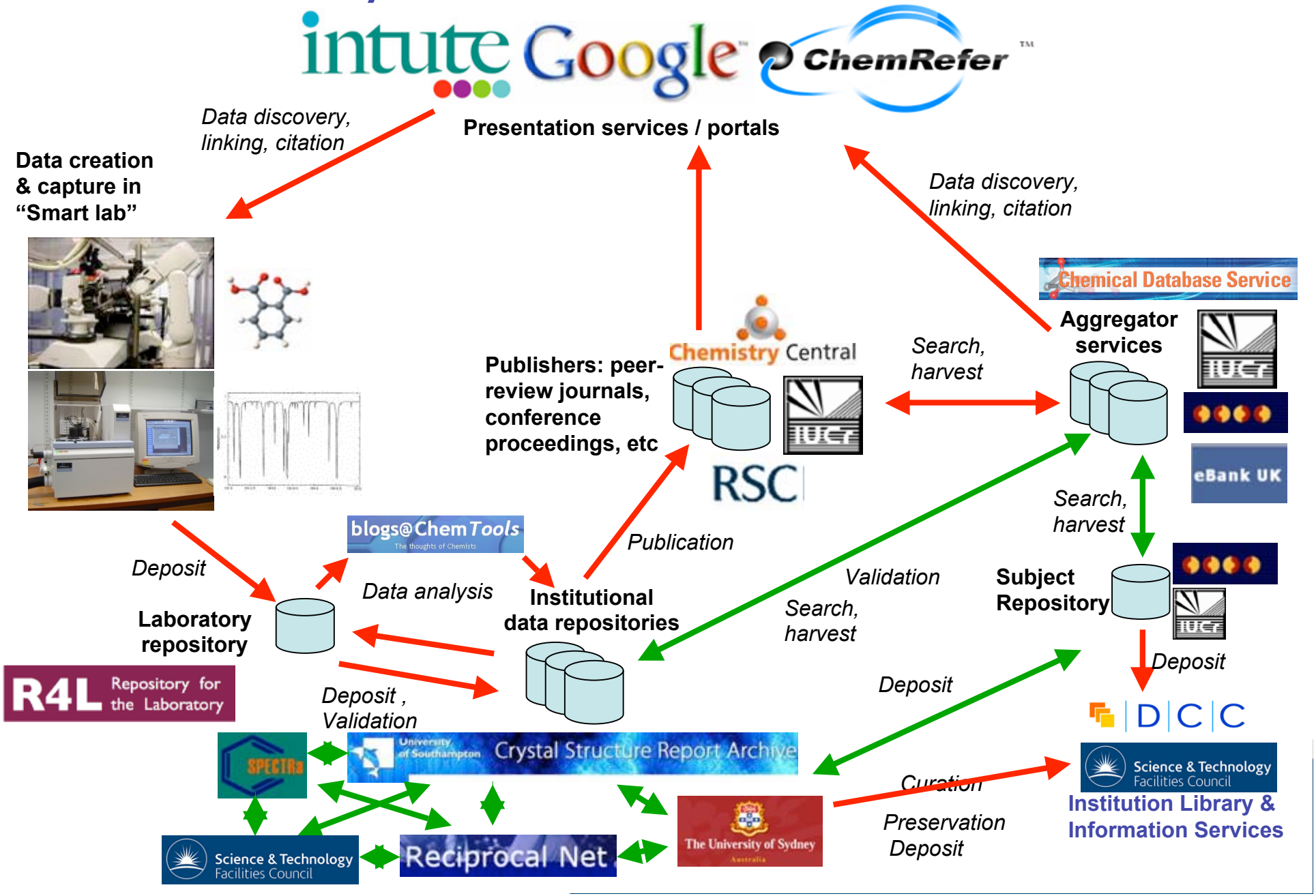
- Across different research lifecycles
- Across institutions
- Across information objects
- Across disciplines
- Across time

Characteristics

- Loosely coupled
- Across different authorities
- Different internal models

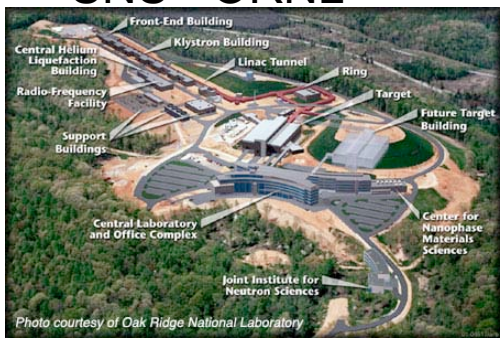
Infrastructure to support science across disciplines, scientific institutions and research groups

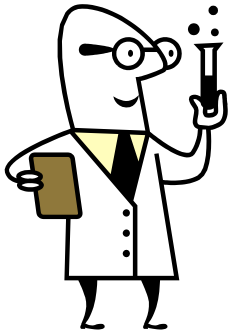
eCrystals 'Data Federation'



e-Infrastructure – Access to Multiple Facilities

SNS - ORNL





CSL - Canada

Data Portal

ISIS – TS1 + 2

DLS

SRS + ERLP

CLF

Science & Technology Facilities Council
e-Science

EDNP

European Data Infrastructure for Neutron and Photon Sources

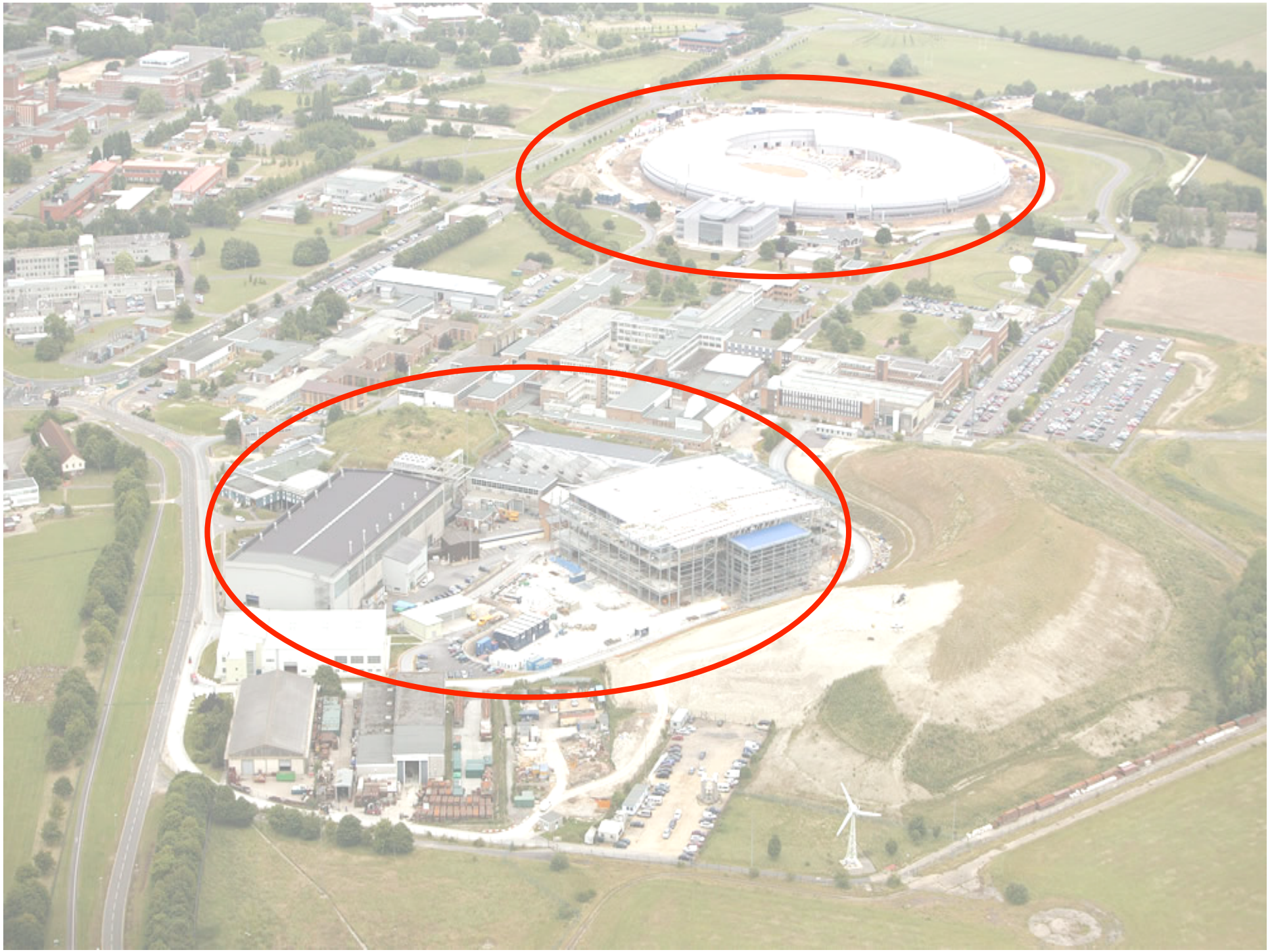
Combining European Neutron and Synchrotron Facilities

Already a common user community

Across many disciplines

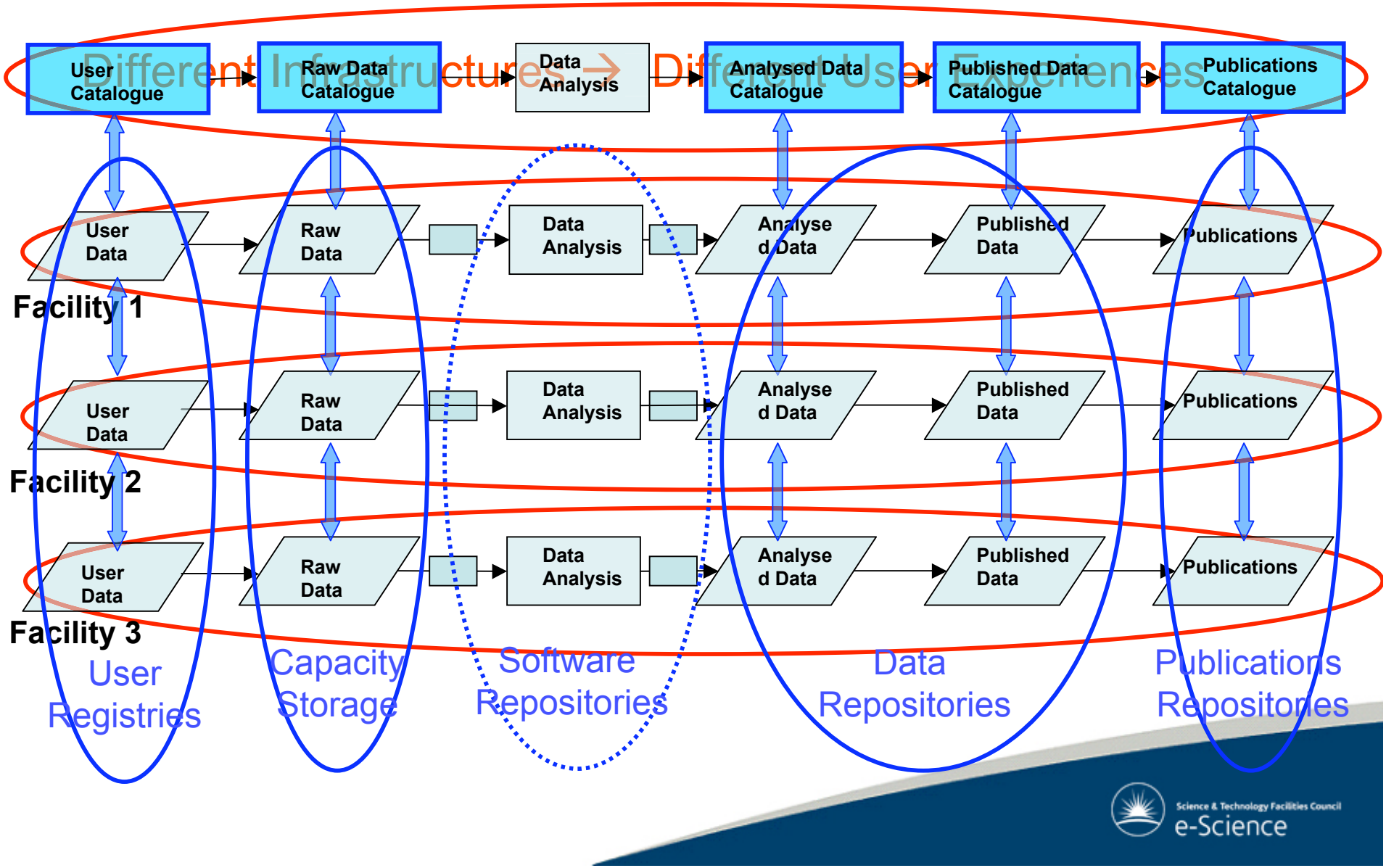
- Materials, chemistry, proteomics, pharmaceuticals, nuclear physics, archaeology …

Science & Technology Facilities Council
e-Science



Integration and interoperation across facilities

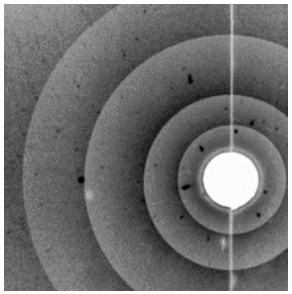
Single Infrastructure → Single User Experience



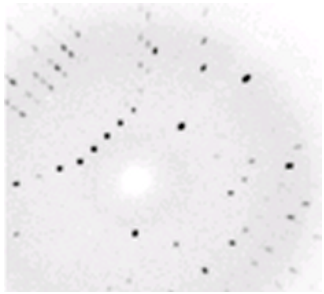
Enabling better science

SCIENCE MASHUPS

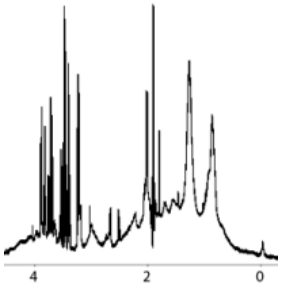
Neutron diffraction

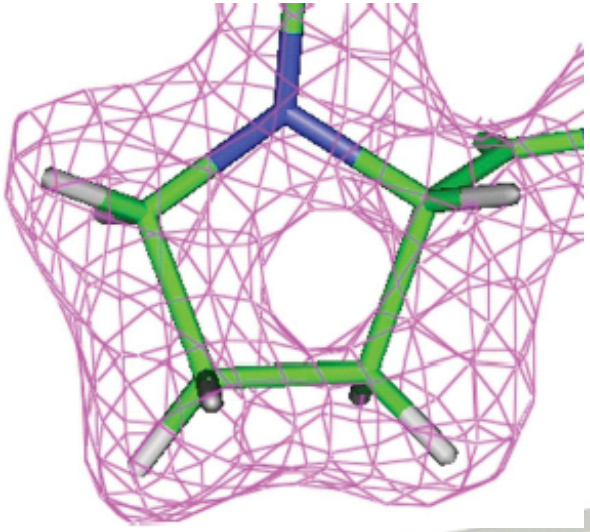


X-ray diffraction



NMR





High-quality structure refinement

Science & Technology Facilities Council
e-Science

Potential Impact

Most of Research Lifecycle

- User Management, Data Collection, Analysis, Publication

- Establish a Production service
  - benefit to users – usability, findability: user info, data, pubs, software
  - benefit to facilities – manageability: users, data, pubs, software

- Outreach and expansion
  - Linking with other facilities in Europe and the wider world
    - USA, Canada, Australia
  - Linking with User communities

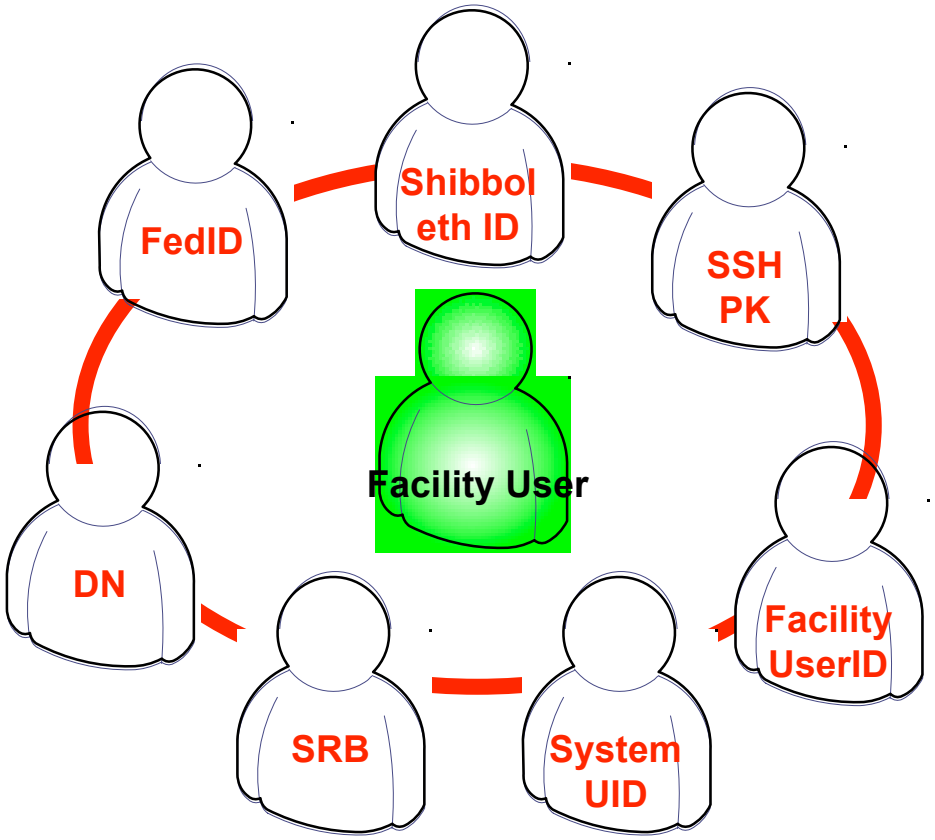
But at the moment, we are still in the planning and discussion phase

Sharing Users

Sharing knowledge of users
- Enhancing level of support for users
- Can correlate similar applications put into different facilities
- Facilities can provide a continuity of service
- Facilities can increase accuracy

Common Authentication
- Common UID ?
- Shibboleth
- Grid Certificates
- SSO at STFC, ShibGrid
- Virtual Organisation Support

Policy Issues
- Data protection
- Institutional Security policy



Sharing Data

Sharing data is hard:

- Different data formats
- Different access rights
- Complex objects
- Maintaining context

Metadata is key

- Structural Metadata (CSMD)
- Conceptual structures (Ontologies) – maintain meaning
- Metadata is hard to collect

Consistent data policies are needed

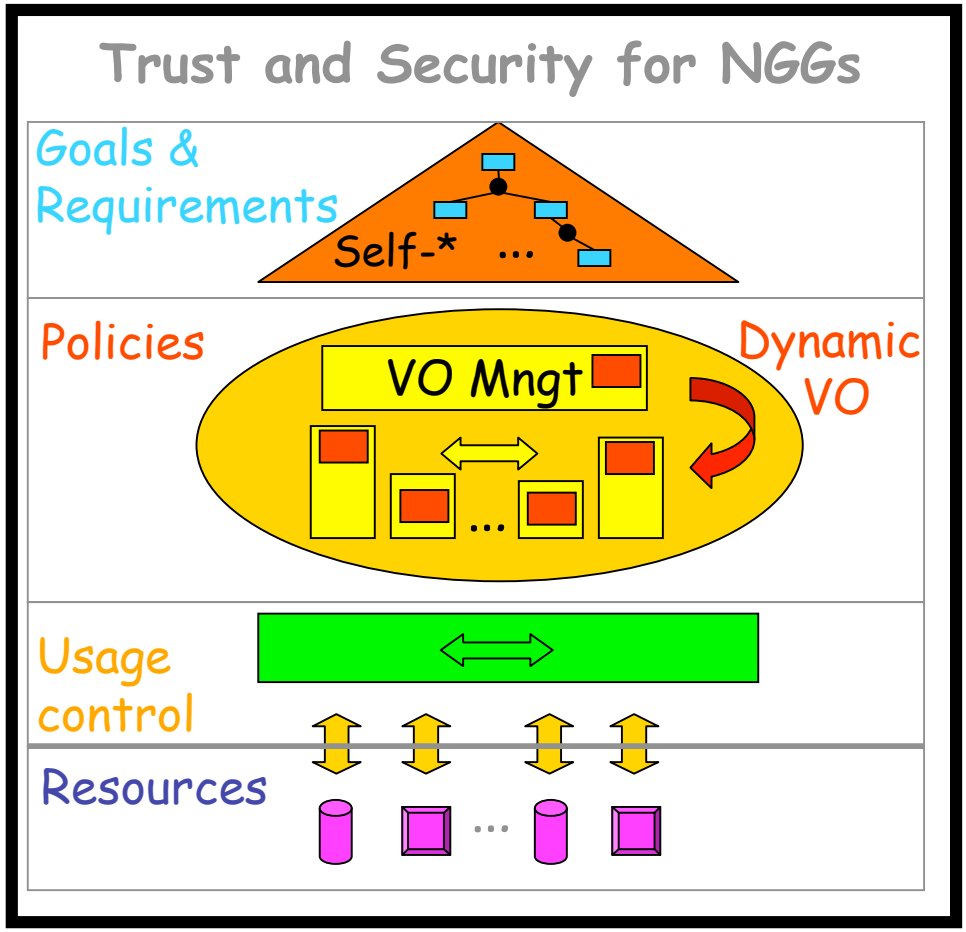
Data Policy

Data policy
- Retention
- Quality
- Access

Learning how to manage policy as part of the SOA infrastructure
- E.g GridTrust
- Consequence – looking at Data Policy

Remains as a very large **Business** question

Use the CSMD to capture the data policy (conditions)



Sharing Software

Analysis software tends to be specialised
- Dependent on specific data formats
- Dependent of nature of data
- Dependent on the particular result to demonstrated

Nevertheless common s/w repositories exist
- GAMS, StarLink, NAG, CCPForge etc

Advantages in sharing it
- Saves programmer effort
- Verification of results
- Common algorithms
- Visualisation tools

Little work on systematic preservation of s/w
- Signficant properties of s/w

Summary

CSMD developed to capture large-facilities science

Successfully deployed in the ICAT/DataPortal Infrastructure

Want to leverage to speed up the science lifecycle from interoperability

Access to resources across institutions and disciplines

Metadata Key Policy Key

Science & Technology Facilities Council
e-Science

Questions?

brian.matthews@stfc.ac.uk

Science & Technology Facilities Council
e-Science